Appendix of “T-LoHo: A Bayesian Regularization Model for Structured Sparsity and Smoothness on Graphs”

Changwoo J. Lee  
Department of Statistics  
Texas A&M University  
c.lee@stat.tamu.edu

Zhao Tang Luo  
Department of Statistics  
Texas A&M University  
ztluo@stat.tamu.edu

Huiyan Sang  
Department of Statistics  
Texas A&M University  
huiyan@stat.tamu.edu

The appendix consists of (1) the details of Algorithm 1 for Bayesian posterior inference in Section A1; (2) the proofs of Propositions 1 and 2 and Theorem 1 in Sections A2, A3, and A4, respectively; (3) additional simulation studies to investigate the hyper-parameter sensitivity analysis, selection criteria, and additional simulation results in Section A5.

A1 Algorithm 1 Details

Algorithm 1 One full iteration of RJMCMC posterior sampler

Step 1. Update $\Pi, K, F$ using the collapsed conditional $[\Pi, K, F|\Lambda, \tau, y]$ where $\beta, \sigma^2$ are integrated out. With probabilities $(p_a, p_b, p_c, p_d)$ summing up to 1, perform one of the following substeps:

1-a. (split) Propose $(\Pi^*, K^* = K+1)$ compatible with $F$, and accept with probability $\min\{1, A_a \cdot P_a \cdot L_a\}$, where $A_a$ is the proposal ratio, $P_a$ is the proposal ratio, $L_a$ is the likelihood ratio.

1-b. (merge) Propose $(\Pi^*, K^* = K-1)$ compatible with $F$, and accept w.p. $\min\{1, A_b \cdot P_b \cdot L_b\}$.

1-c. (change) Propose $(\Pi^*, K^* = K)$ compatible with $F$, and accept w.p. $\min\{1, A_c \cdot P_c \cdot L_c\}$.

1-d. (hyper) Update $F^*$ compatible with current $\Pi$.

Step 2. Jointly update $(\tau, \sigma^2, \beta)$ from $[\tau, \sigma^2, \beta | \Lambda, \Pi, K, F, y]$, by performing:

2-1. Update $\tau$ from $[\tau | \Lambda, \Pi, K, F, y]$, using a Metropolis-Hastings sampler.

2-2. Update $\sigma^2$ from $[\sigma^2 | \tau, \Lambda, \Pi, K, F, y]$ with an inverse gamma distribution.

2-3. Update $\beta$ from $[\beta | \sigma^2, \tau, \Lambda, \Pi, K, F, y]$ with a multivariate normal distribution.

Step 3. Update $\Lambda$ from $[\Lambda | \tau, \sigma^2, \beta, \Pi, K, F, y]$ using a slice sampler.

A1.1 Step 1 Details

Step 1 is motivated by Algorithm 1 of Luo et al. (2021). It updates parameters $\Pi, K$ and $F$ which determine the number of cluster and its shape. Note that step 1-d does not directly update $(\Pi, K)$ but updates the underlying MSF $F$ which allows the sampler to explore different candidates of $(\Pi, K)$. Here the values of $p_a, p_b, p_c$ depend on $K$, and $p_d$ is chosen to be small enough such as 0.05 to give the sampler enough time to explore $(\Pi, K)$ compatible with current $F$.

Proposal probabilities for each move $(p_a, p_b, p_c, p_d)$ depend on the current number of cluster $K$. When $n_c < K < p$, we set $(p_a, p_b, p_c, p_d) = (0.425, 0.425, 0.1, 0.05)$. For the boundary cases when $K = n_c$ and $K = p$, we set $(p_a, p_b, p_c, p_d) = (0.95, 0, 0, 0.05)$ and $(p_a, p_b, p_c, p_d) = (0, 0.95, 0, 0.05)$ respectively.

By proposition 1, there are two types of edges in $F = (V, E^F)$ that determine $\Pi$: a set of cut edges (between-cluster edge) $E^C$ and a set of within-cluster edges $E^F \cap E^C$. For step 1-a (split) with
We propose a new proposal $\Pi^*$ by randomly choosing one edge from the within-cluster edge set $E^F \setminus E^C$ with equal probability and switch it to a cut-edge. For step 1-b (merge) with $K^* = K - 1$, we propose a new proposal $\Pi^*$ by randomly choosing one edge from the cut edge set $E^C$ with equal probability and switch it to a within-cluster edge. For step 1-c (change), $\Pi^*$ is proposed by successively performing split and merge steps. Care is needed regarding local shrinkage parameters $\{\lambda_k\}_{k=1}^K$, since a birth/death of new cluster accompanies adding/deleting a local shrinkage parameter $\lambda^*$. For step 1-a (split), let $\lambda^0$ be a splitting parameter of the splitting cluster $C^0$. By selecting a new cut-edge, we divide $C^0_i \cup C^0_j$ such that $|C^0_i| \geq |C^0_j|$ (if two cluster sizes are the same, then based on a proposed cut-edge $(i, j)$ with $i < j$, $C^0_i$ is the cluster that includes $i$). Then the local parameter of the bigger cluster $C^0_i$ is inherited from the original cluster $\lambda^0_i$, and the local parameter of the smaller cluster is drawn from its prior $\lambda^0_j \sim \mathcal{C}^0(0, 1)$. Similarly, for step 1-b (merge), we can define $C^0 = C^0_i \cup C^0_j$ such that $|C^0_i| \geq |C^0_j|$ and let $\lambda^0 = \lambda^0_i$. This type of proposal makes step 1 satisfy detailed balance condition (Green, 1995).

The collapsed conditional $[\Pi, K, \mathcal{F}|\Lambda, \tau, \Psi]$ (where $\tilde{\beta}$ and $\sigma^2$ are integrated out) is proportional to

$$[\Pi, K, \mathcal{F}|\Lambda, \tau, \Psi] \propto \int \int \mathcal{N}(\Psi|X, \sigma^2 I_n) \mathcal{N}(\tilde{\beta}|0, \sigma^2 \tau^2 \Lambda) \sigma^{-2} d\tilde{\beta} d\sigma^2 \left(\frac{p - n_c}{K - n_c}\right)^{-1} (1 - c)^K$$

(A1)

$$\propto \int \mathcal{N}(\Psi|0, \sigma^2 (I_n + \tau^2 XX^T)) \times \sigma^{-2} d\sigma^2 \times \left(\frac{p - n_c}{K - n_c}\right)^{-1} (1 - c)^K$$

(A2)

$$\propto |\Sigma|^{-1/2} (y^\top \Sigma^{-1} y/2)^{-n/2} \times \left(\frac{p - n_c}{K - n_c}\right)^{-1} (1 - c)^K$$

(A3)

where $\Sigma = I_n + \tau^2 XX^T$. Note that $X = \Phi \beta^T$ is a function of $\Pi$ since $\Pi$ determines $\Phi$. Line (A3) can be decomposed into a likelihood part $|\Sigma|^{-1/2} (y^\top \Sigma^{-1} y/2)^{-n/2}$ and a prior part $\left(\frac{p - n_c}{K - n_c}\right)^{-1} (1 - c)^K$. Then the split step proposal $(\Pi^*, K^* = K + 1)$ is accepted with probability

$$\min \left\{ \frac{1}{A_a \cdot P_a \cdot L_a} \right\}$$

Similarly, the merge step proposal $(\Pi^*, K^* = K - 1)$ is accepted with probability

$$\min \left\{ \frac{1}{A_b \cdot P_b \cdot L_b} \right\}$$

and the change step proposal $(\Pi^*, K^* = K)$ is accepted with probability

$$\min \left\{ \frac{1}{A_c \cdot P_c \cdot L_c} \right\}$$

For step 1-d (hyper), conditioning on the current partition, $\Pi$ divides the edge set $E$ of $G$ into the between-cluster edge set $E^b$ and within-cluster edge set $E^w$. We sample $F$ by sampling $W_{ij} \sim \text{Unif}(0, 1/2)$ for $(i, j) \in E^w$, $W_{ij} \sim \text{Unif}(1/2, 1)$ for $(i, j) \in E^b$, and letting $F = \text{MSF}(G, W)$. $2$
To see this is a correct sampler, note that $p(W) \propto 1$ is symmetric and $F = MSF(G, W)$ depends only through the order of random edge weights $\{W_{ij}\}_{(i,j) \in E}$; see Luo et al. (2021) for detailed discussion.

As briefly discussed in Section 3.1, we apply several computational strategies to efficiently calculate $L$. Plus, if $X = I_n$, we can marginalize out $\Lambda$ when calculating $L$ to improve MCMC mixing and avoid local parameter proposals $\lambda^*$; see Section A1.3.

### A1.2 Step 2 and Step 3 Details

Steps 2 and 3 follow a similar procedure of Johndrow et al. (2020) as detailed below.

1. **Step 2-1:** Propose $\log(\tau^*) \sim \mathcal{N}(\log(\tau), s)$, accept $\tau^*$ w.p. $p(\cdot) - \frac{|\Sigma^*|^{-1/2}(y^\top y - n/2) + \tau^2/2}{1 + \tau^2/2\sigma^2} \frac{\tau^*}{\tau^*}$.

2. **Step 2-2:** Sample $\beta | \sigma^2, \tau, \Lambda, y \sim \mathcal{N}_K \left((\tau^{-2}\Lambda^{-1} + \bar{X}^\top \bar{X})^{-1}\bar{X}^\top y, \sigma^2(\tau^{-2}\Lambda^{-1} + \bar{X}^\top \bar{X})^{-1}\right)$.

3. **Step 2-3:** Sample $\lambda_k \sim p(\lambda_k | \tau, \sigma^2, \beta) \propto \frac{1}{1 + \lambda_k^2} \cdot \frac{1}{1 + \lambda_k^2} \exp \left(-\frac{\lambda_k^2}{2\sigma^2 + \tau^2_2}\right)$, $k = 1, \ldots, K$.

In step (2-1), the Metropolis-Hastings proposal variance $s$ is adjusted every 1000th iteration to maintain acceptance ratio close to 0.35 (Haario et al., 2001). In step (2-3), Rue (2001)’s algorithm is used and we’ll explain its benefits in the next subsection. In step (3), we use slice sampler (Neal, 2003; Polson et al., 2014) by reparametrizing $\eta_k := 1/\lambda_k^2$ and introducing auxiliary variables $\{u_k\}_{K=1}$.

Specifically, first sample $u_k \sim \text{Unif}(0, 1/(1 + \eta_k))$ and then sample $\eta_k \sim \text{Exp}(2\tau_2^2/\beta^2_2) 1(0 < \eta_k < 1/u_k - 1)$, $k = 1, \ldots, K$ where $\text{Exp}(\theta)|0 < \eta < a)$ is a truncated exponential distribution with parameter $\theta$ (mean before truncation) and support $(0, a)$.

### A1.3 Computational Strategies

When calculating the likelihood, we utilize Sherman-Woodbury-Morrison formula, matrix determinant lemma and Cholesky decomposition for the efficient calculation of $y^\top \Sigma^{-1} y$ and $|\Sigma|$. If we let $R$ be a $K \times K$ right triangular Cholesky factor such that $R^\top R = \tau^{-2}\Lambda^{-1} + \bar{X}^\top \bar{X}$, then

$$y^\top \Sigma^{-1} y = y^\top y - y^\top \bar{X}(\tau^{-2}\Lambda^{-1} + \bar{X}^\top \bar{X})^{-1}\bar{X}^\top y$$

$$|\Sigma| = |\tau^{-2}\Lambda^{-1} + \bar{X}^\top \bar{X}| = |R|^2 \cdot \prod_{k=1}^K (\lambda_k^2 \tau^2)$$

(A4) (A5)

It is evident to see the computational benefits of using Cholesky factor $R$. First in (A4), $y^\top y$ is a constant and $y^\top \bar{X}(R^\top R)^{-1}\bar{X}^\top y = \|v\|_2^2$ where $v$ is obtained by solving $R^\top v = \bar{X}^\top y$ using only forward substitution. In (A5), $|R|$ is simply a product of its diagonal elements. Next in step (2-3), using Rue (2001)’s algorithm we draw $\beta$ by back substitution $\sigma^{-1}R\beta = (\sigma^{-1}v + z)$ where $z \sim \mathcal{N}_K(0, I_K)$.

During the MCMC, the Cholesky factor $R$ should be updated whenever a change of $\bar{X}$, $\tau$ or $\Lambda$ takes place. For instance, in step 1 we have a new $\bar{X}^* = X\Phi^*$ which leads to $R^*R^* = \tau^{-2}\Lambda^{-1} + \bar{X}^*\bar{X}^*$ where a row and column are added/deleted in step 1-a/1-b respectively. A naive calculation of $R$ requires $O(nK^2)$ (assuming $K < n$), but we avoid this by directly updating $R$ from its previous value. Specifically, in step (1), $X$ (and $\Lambda$) changes, and we use the rank-1 update/downdate of Cholesky factor (Golub and Van Loan, 2013, Sec. 6.5.4), (Osborne, 2010, Appendix B) which reduces the cost to $O(K^3)$. In steps (2-1) and (3), the diagonal part of $\tau^{-2}\Lambda^{-1} + \bar{X}^\top \bar{X}$ changes, and we propose a tailored updating algorithm, *Cholesky diagonal update*, and presented it in Algorithm 2. It involves about $(2/3)K^3$ FLOPs when $R$ is $K \times K$. Thus, we can directly update $R$ with the diagonal modification vector $v = (\tau^{-2} - \tau^{-2})\text{diag}(\Lambda)^{-1}$ for step (2-1) and $v = \tau^{-2}\text{diag}(\Lambda^*)^{-1} - \text{diag}(\Lambda)^{-1}$ for step 3. Combined with forward/backward substitution and Rue (2001)’s algorithm described above, these Cholesky updating strategies speed up the algorithm significantly, especially when $K$ is small.
Algorithm 2 Cholesky diagonal update

Input: Right triangular Cholesky factor \( \text{chol}(A) = R \in \mathbb{R}^{n \times n} \), diagonal part modification vector \( v \in \mathbb{R}^n \). Here \( A \) and \( A + \text{diag}(v) \) are both assumed to be positive definite.

Output: Right triangular Cholesky factor \( U = \text{chol}(A + \text{diag}(v)) \).

1: Initialize \( U \in \mathbb{R}^{n \times n} \) with zero elements.
2: \( w \leftarrow \text{colSums}(R \circ R) + v \) where \( \circ \) indicates elementwise product.
3: for \( i \) in 1 to \( n \) do
4: \( U[i, i] \leftarrow \sqrt{w[i]} - \sum_{k=1}^{i-1} U[k, i]^2 \)
5: \( s \leftarrow \sum_{k=1}^{i-1} (R[k, i]R[k, (i + 1) : n] - U[k, i]U[k, (i + 1) : n]) \)
6: \( U[i, (i + 1) : n] \leftarrow (R[i, i]R[i, (i + 1) : n] + s)/U[i, i] \)
7: end for
8: return \( U = \text{chol}(A + \text{diag}(v)) \)

In addition, if \( X = I_n \) we can marginalize out \( \Lambda \) instead of \( \sigma^2 \) in step 1 which leads to improved mixing of MCMC chain and avoiding \( \Lambda^* \) proposal procedure, that is, we can write a new likelihood ratio \( \mathcal{L}' \) as

\[
\mathcal{L}' = \frac{\int \cdots \int p(\lambda, \Lambda, \tau, \sigma^2) \prod_{k=1}^{K^*} p(\lambda_k) d\lambda_k}{\int \cdots \int p(\lambda, \Lambda, \tau, \sigma^2) \prod_{k=1}^{K} p(\lambda_k) d\lambda_k},
\]

where \( \mathcal{L}' \) is the likelihood ratio of the update step. Similarly, we can calculate the likelihood ratio of the merge step \( \mathcal{L}_c \) by inverting the expression (A7), and the change step \( \mathcal{L}_c \) is the product of the two. Here the product of the prior and proposal ratios remains the same as before; i.e., \( A' \cdot P' = A \cdot P \) for steps (1-a), (1-b), and (1-c). Thus when \( X = I_n \), we can update \( (\Pi, K) \) using collapsed conditional \( [\Pi, K; \sigma^2, \tau, y] \) where all local parameters \( (\beta_k, \lambda_k) \) are integrated out, with acceptance ratio \( \alpha' = \min(1, A' \cdot P' \cdot \mathcal{L}') \). Computation of other parts in algorithm may also be simplified, since \( X = I_n \) implies \( X^\top X = I_K \) and thus \( R = \text{diag}(\tau^{-2} \lambda_{\text{chol}}^{-2} + 1, \cdots, \tau^{-2} \lambda_{\text{chol}}^{-2} + 1) \).

A2 Proof of Proposition 1

Proof. Let \( G_j = (V_j, E_j) \), \( j = 1, \ldots, n_c \) be connected components of \( G \). Then any graph partition \( \Pi = \{C_1, \ldots, C_K\} \) of \( G \), where \( K \geq n_c \), can be divided into \( n_c \) disjoint subsets \( \Pi = \bigcup_{j=1}^{n_c} \Pi_j \) such that each \( \Pi_j \) corresponds to a graph partition of \( G_j \), according to the definition of contiguous graph partitions. Note each \( G_j \) is a connected graph. By applying Proposition 2 of Luo et al. (2021) for each \( G_j \), there exists a spanning tree \( T_j = (V_j, E_j) \) with \( |E_j| = |V_j| - 1 \), and a set of cut-edges \( E_j^C \subset E_j \) with \( |E_j^C| = |\Pi_j| - 1 \) such that \( T_j \) and \( E_j^C \) induces \( \Pi_j \). Then we can construct a spanning forest \( F = (V, E^F) \) where \( E^F = \bigcup_{j=1}^{n_c} E_j \) and a set of cut-edges \( E^C = \bigcup_{j=1}^{n_c} E_j^C \) with \( |E^C| = \sum_{j=1}^{n_c} (|\Pi_j| - 1) = K - n_c \) which completes the proof. \( \square \)
A3 Proof of Proposition 2 and Median Variance Matching Criterion

First we present the proof of Proposition 2.

Proof. Here $\sigma^2, \tau_1, \tau_2$ are assumed to be fixed. If $\mu_1 | \lambda_1 \sim \mathcal{N}(0, \lambda_1^2 \tau_1^2 \sigma^2)$ and $\mu_2 | \lambda_2 \sim \mathcal{N}(0, \lambda_2^2 \tau_2^2 \sigma^2)$ independently, the conditional distribution of $\delta = (\mu_1 - \mu_2) / \sigma$ given $\lambda_1, \lambda_2$ is

$$\delta | \lambda_1, \lambda_2 \sim \mathcal{N}(0, \lambda_1^2 \tau_1^2 + \lambda_2^2 \tau_2^2).$$

When $\lambda \sim \mathcal{C}^+(0, 1)$ independently, it is easy to see that density of $V = \tau^2 \lambda^2$ is given by $f_V(v|\tau) = \tau / (\tau \sqrt{\pi(v + \tau^2)})$, $v > 0$. Then since $\lambda_1^2 \tau_1^2$ and $\lambda_2^2 \tau_2^2$ are independent, the marginal distribution of $W := \lambda_1^2 \tau_1^2 + \lambda_2^2 \tau_2^2$ can be calculated by convolution formula; i.e.,

$$f_W(w) = \int_{-\infty}^{\infty} f_V(w - v|\tau_1) f_V(v|\tau_2) dv = \frac{2}{\pi^2(w + \tau_1^2 + \tau_2^2)} \left[ \frac{\tau_2 \tan^{-1}\left(\frac{\tau_2 \sqrt{w + \tau_2^2}}{\sqrt{w + \tau_1^2}}\right)}{\sqrt{w + \tau_1^2}} + \frac{\tau_1 \tan^{-1}\left(\frac{\tau_1 \sqrt{w + \tau_1^2}}{\sqrt{w + \tau_2^2}}\right)}{\sqrt{w + \tau_2^2}} \right]_{v=0}^{v=w}$$

Therefore the marginal distribution of $\delta$ can be written as

$$\pi_\Delta(\delta|\tau_1, \tau_2) = \int_0^{\infty} \mathcal{N}(\delta|0, w) f_W(w|\tau_1, \tau_2) dw$$

Now we explain the median variance matching criterion under the different settings of $(n_1, n_2)$. The cumulative distribution function of $W$ is

$$F_W(w) = \frac{2}{\pi} \left[ \tan^{-1}\left(\frac{\sqrt{w + \tau_2^2}}{\tau_1}\right) + \tan^{-1}\left(\frac{\sqrt{w + \tau_1^2}}{\tau_2}\right) \right] - 1$$

for $w \geq 0$ and 0 otherwise. By solving $F_W(w) = 1/2$, the median is

$$\text{median}(W) = \tau_1^2 + \tau_2^2 + 2\tau_1 \tau_2 \sqrt{2}$$

Since moments of $W$ do not exist, the closed form of median gives a useful measure of ‘center’ of $W$.

Following Section 3.2, consider the normal means model $X = I_n$ with two clusters $C_1, C_2$ with sizes $n_1, n_2$. Then T-LoHo induces a prior on the two cluster mean $\mu_1$ and $\mu_2$ as

$$\mu_1 | \lambda_1, \tau, \sigma^2 \sim \mathcal{N}\left(0, \frac{1}{n_1} \lambda_1^2 \tau^2 \sigma^2\right), \quad \mu_2 | \lambda_2, \tau, \sigma^2 \sim \mathcal{N}\left(0, \frac{1}{n_2} \lambda_2^2 \tau^2 \sigma^2\right)$$

Thus when $\lambda_1, \lambda_2 \stackrel{iid}{\sim} \mathcal{C}^+(0, 1)$ are integrated out, the induced prior on $\delta = (\mu_1 - \mu_2) / \sigma$ is

$$\delta \sim \pi_\Delta(\delta|\tau_1 = \tau / \sqrt{n_1}, \tau_2 = \tau / \sqrt{n_2})$$

We match the median of $W$ as 1 for a fair comparison with $\mathcal{N}(0, 1)$. When $n_1 = n_2 = n$ (balanced), we set $\tau = \sqrt{n} / \sqrt{2 + 2\sqrt{2}}$. When $n_1 : n_2 = 9 : 1$ (unbalanced), we set $\tau = \sqrt{n_1} / \sqrt{10 + 6\sqrt{2}}$.

A4 Proof of Theorem 1

Let $(\beta^*, \hat{\beta}^*, \sigma^*)$ denote the true $\beta, \hat{\beta}$ and $\sigma$ respectively. Let $\xi^* = \{j \in V : \beta^*_j \neq 0\}$ denote the true active set of indices. Let $\hat{\Pi}$ denote an arbitrary partition of $V = \{1, \cdots, p\}$ whose corresponding partition of $\xi^*$ is determined by removing those edges with $\beta^*_i - \hat{\beta}^*_i > 0$ from the subgraph of any
We also introduce the following notations. Let \( f \) denote all unique partitions that have at most \( g^*_c \) clusters and their corresponding partitions of \( \xi^* \) are nested in the true partition of \( \xi^* \) for some constant \( c_3 > 0 \).

Below, we consider the case when \( p \) is much larger than \( n \) and establish posterior concentration results for the T-LoHo model as \( n \) goes to infinity. Our results rely on the following assumptions:

(A-1) The graph satisfies \( g^*_c < n / \log p, n_c = o(g^*_c), \) and \( \log |P_n| = O(g^*_c \log p). \)

(A-2) All the covariates are uniformly bounded. There exist some fixed constant \( \lambda_0 > 0 \), such that \( \lambda_{\min}(X^T X) \geq \lambda_0 \) for any partition in \( P_n \).

(A-3) \( \max_j |\hat{\beta}_j|/\sigma_j < L, \) where \( \log(L) = O(\log p). \)

(A-4) \( -\log \tau \leq \log p, \tau < p^{-(2+\epsilon_\tau)} \sqrt{g^*_c \log p/n}, 1 - \epsilon \geq p^{-c_\epsilon}, \) and \( \min_{\alpha,\beta \in \{a^2, \sigma^2 (1+c_\alpha c_\beta)^2\}} \pi(\sigma^2) > 0 \) for some positive constants \( c_\epsilon, c_\alpha, c_\beta. \)

**Theorem 1.** (Posterior contraction) Under Assumptions (A-1) to (A-4), there exists a large enough constant \( M_1 > 0 \) and \( \varepsilon_j = \sqrt{g^*_c \log p/n} \) such that the posterior distribution satisfies \( \pi_n(\|\beta - \beta^*\|^2_2 \geq M_1 \sigma^2 \varepsilon_j | y) \leq \exp(-c_1 n \varepsilon_j^2) \) with probability \( 1 - \exp(-c_2 n \varepsilon_j^2) \) for some constants \( c_1 > 0 \) and \( c_2 > 0 \).

**A4.1 Notations and Lemmas**

We begin by introducing some asymptotic notations. Given two positive sequences \( \{a_n\} \) and \( \{b_n\}, a_n \geq b_n \) means \( \lim_{n \to \infty} (a_n/b_n) = \infty \) and \( a_n \leq b_n \) means \( 0 < \liminf_{n \to \infty} (a_n/b_n) \leq \limsup_{n \to \infty} (a_n/b_n) < \infty. \) We denote the \( \ell_2 \) norm by \( \| \cdot \|. \) We use \( p \) to denote the dimension of \( \beta. \)

We also introduce the following notations. Let \( f_\theta \) be the likelihood function of parameter \( \theta \in \Theta_n \) given \( y \) from a data generation model, whose prior is \( \pi(\theta) \) and true parameter value is \( \theta^0. \) We use \( f^* = f_{\theta^0} \) to denote the density function of \( y \) when \( \theta = \theta^0, m(y) \) to denote the marginal density of \( y, E_\theta, E^* \) to denote the expectations under \( f_\theta \) and \( f^* \) respectively, \( \Pr^* \) to denote the probability measure under \( f^* \), and \( \pi_n \) to denote the posterior distribution given \( y. \)

We let \( \Pi(\beta) \) denote the contiguous partition of \( \{1, \cdots, p\} \) with respect to \( G \) according to \( \beta. \) Define \( \xi^1(\beta, \sigma) := \bigcup_{\{c_1, \cdots, c_\beta \}} \chi_{c_1, \cdots, c_\beta} \) such that it only includes the set of indices in \( \{1, \cdots, p\} \) whose corresponding \( |\hat{\beta}|/\sigma \) is greater than \( \varepsilon_n/p, \) and \( \Pi^1(\beta, \sigma) \) be the partition of \( \xi^1(\beta, \sigma) \) under \( \Pi(\beta). \)

Below, we also use \( \Pi(\xi) \) to denote the partition of set \( \xi(\beta, \sigma) \) when there is no risk of confusion.

We now state some lemmas that will be used in the proof of Theorem 1.

**Lemma 1.** (Lemma 1 of Laurent and Massart (2000)) Let \( \chi^2_d \) be a chi-square distribution with degree of freedom \( d. \) Then the following concentration inequalities hold for any \( x > 0: \)

\[
\Pr \left( \chi^2_d > d + 2x + 2\sqrt{dx} \right) \leq \exp(-x)
\]

and

\[
\Pr \left( \chi^2_d < d - 2\sqrt{dx} \right) \leq \exp(-x).
\]

**Lemma 2.** (Lemmas 10 and 11 of Banerjee (2021)) Let \( \pi_{HS}(\beta | \sigma = 1, \tau) \) denote the horseshoe prior density on \( \beta \) assuming a fixed \( \tau > 0 \) and \( \sigma^2 = 1. \) Under Assumptions (A-3) and (A-4), for some \( c_\tau > 0, \) we have

\[
\int_{-\varepsilon_n/p}^{\varepsilon_n/p} \pi_{HS}(\beta; 1, \tau) d\beta \geq 1 - p^{-(1+c_\tau)}, -\log \left( \inf_{\beta \in [-L, L]} \pi_{HS}(\beta; 1, \tau) \right) = O(\log p) \tag{A8}
\]

**Lemma 3.** (Lemma A.3 of Song and Cheng (2020)) Let \( B_n \) and \( C_n \) be two subsets of the parameter space \( \Theta_n, \) and \( \phi_n \) be a test function satisfying \( \phi_n(D_n) \in [0, 1] \) for any data \( D_n. \) If \( \pi(B_n) \leq b_n, E^* \{\phi_n(D_n)\} \leq b'_n, \sup_{\theta \in \Theta_n} E \{1 - \phi_n(D_n)\} \leq c_n, \) and

\[
\Pr^* \left( m(D_n) \geq a_n \right) \geq 1 - a_n'.
\]

Then

\[
E^* \{\pi_n(C_n) \cup B_n | D_n) \} \leq \frac{b_n + c_n}{a_n} + a_n' + b'_n.
\]
A4.2 Proof of Theorem 1

Proof. The proof of Theorem 1 proceeds by verifying the conditions in Lemma 3 in three parts, adapting the proof in Song and Liang (2017) originally used for a Bayesian variable selection problem. The first part involves the construction of $B_n$ and $C_n$, the second part shows the existence of testing functions $\phi_n$, and the last part proves the evidence lower bound.

Part 1 (Sieve construction): We define

$$C_n = \left\{ (\beta, \sigma) : \| \beta - \beta^* \| \leq M_1 \sigma^* \varepsilon_n \text{ and } \frac{1 - \varepsilon_n}{1 + \varepsilon_n} < \frac{\sigma^2}{\sigma^2 + 2} < \frac{1 + \varepsilon_n}{1 - \varepsilon_n} \right\} \setminus B_n,$$

and

$$B_n = \left\{ (\beta, \sigma) : \Pi^1(\beta, \sigma) \text{ has at least } c_3 g_n^* \text{ clusters } \right\}.$$

for some constant $c_3 > 0$.

Recall $K \equiv |\Pi(\beta, \sigma)|$ follows a left truncated geometric distribution. Conditional on $K$, the number of clusters whose corresponding $|\beta/\sigma|$ exceeds $\varepsilon_n/p$ follows a Binomial $(K, v_n)$ distribution, where

$$v_n = \int_{|x| \geq \varepsilon_n/p} \pi_{h_0}(x; 1, \tau) \, dx$$

and $\pi_{h_0}(x)$ is the horseshoe prior density function for $\beta/\sigma$. From Lemma 2, it is known that $v_n = O(p n^{-(1+\gamma)}).$ Following a similar proof as in Song and Liang (2017), we can show that for any $K > c_3 g_n^*$, $\Pr(\text{Binomial}(K, v_n) \geq c_3 g_n^*) \leq \exp(-c_1 n \varepsilon_n^2)$. For some constant $c_1$ using a sharp tail bound for binomial distributions in Theorem 1 of Zubkov and Serov (2013). Conditional on any given spanning forest $F$, $\pi(B_n | F) = \Pr(\text{Binomial}(K, v_n) \geq c_3 g_n^* | K \geq c_3 g_n^*) \Pr(K \geq c_3 g_n^*) \leq \exp(-c_1 n \varepsilon_n^2)$. Note that $c_1$ does not depend on $F$. Therefore there exists some constant $c_1$ such that

$$\pi(B_n) = \sum_F \pi(B_n | F) \pi(F) \leq \exp(-c_1 n \varepsilon_n^2). \quad (A9)$$

Part 2 (Existence of testing function). Given an arbitrary partition $P(\xi)$ of an arbitrary set $\xi$ with $K_{\xi}$ clusters, let $X_{\xi}$ be the corresponding transformed design matrix, and $H_{\xi}$ be the hat matrix of $X_{\xi}$. Define $\hat{\beta}_\xi = (X_{\xi}^T X_{\xi})^{-1} X_{\xi}^T y$, $\hat{\sigma}_\xi^2 = y^T (I - H_{\xi}) y / (n - K_{\xi})$, and $\hat{\beta}_\xi^* = P(\Pi(\xi)) \beta^*(\xi)$.

We define a test function

$$\phi(y) = \max_{\{x \in x^*\}} 1(|\hat{\beta}_\xi - \hat{\beta}_\xi^*| \geq \sigma^* \varepsilon_n \text{ and } |\hat{\sigma}_\xi^2 / \sigma^2 - 1| \geq \varepsilon_n),$$

for some fixed $c_3 > 0$, where $P(\Pi^*(\xi^*))$ is the true partition of $\xi^*$.

From standard linear regression results, using the true parameter ($\beta^*, \sigma^*$) and the restricted eigenvalue assumption in (A-2), we have $\Pr(|\hat{\beta}_\xi - \hat{\beta}_\xi^*| \geq \varepsilon_n) = \Pr \left( \frac{|\chi_{n-K_{\xi}}^2/(n - K_{\xi}) - 1| \geq \varepsilon_n} \right)$ and $\Pr(|\hat{\sigma}_\xi^2 | \sigma^2 - 1| \geq \varepsilon_n) \leq \exp(-c_21 n \varepsilon_n^2)$. From the union bound and the last part of Assumption (A-1), we can now bound the type-I error of the test function as follows

$$\operatorname{E}^* \{\phi(y)\} \leq P_n \cdot \exp(-c_21 n \varepsilon_n^2) \leq \exp(-c_21 n \varepsilon_n^2), \quad (A10)$$

for some constant $c_21 > 0$ and large $n \varepsilon_n^2 / (g_n^* \log p)$.}

Next we bound the type II error part $\sup_{(\beta, \sigma) \in C_n} \operatorname{E} \{1 - \phi(y)\}$. We define two subsets $C_{n(1)}$ and $C_{n(2)}$ such that $C_n \subset C_{n(1)} \cup C_{n(2)}$ and analyze them separately, where

$$C_{n(1)} = \left\{ (\beta, \sigma) : \| \beta - \beta^* \| > M_1 \sigma^* \varepsilon_n, \frac{\sigma^2}{\sigma^2 + 2} < \frac{1 + \varepsilon_n}{1 - \varepsilon_n} \right\} \cap B_n^c,$$

and

$$C_{n(2)} = \left\{ \sigma : \frac{\sigma^2}{\sigma^2 + 2} \leq \frac{1 - \varepsilon_n}{1 + \varepsilon_n} \text{ or } \frac{\sigma^2}{\sigma^2 + 2} \geq \frac{1 + \varepsilon_n}{1 - \varepsilon_n} \right\} \cap B_n^c.$$
For any $\beta \in C_n$, let $F(\beta)$ be any spanning forest that can induce $\Pi(\beta)$. Consider a set $\tilde{\xi}$ formed by keeping those column indices whose $|\beta|/\sigma > \varepsilon_n/p$ or $\beta^* > 0$. Thus $\tilde{\xi} \supset \xi^*$. We then form a partition $\Pi(\tilde{\xi})$ of $\tilde{\xi}$ by removing those edges with $\beta^*_i - \beta^*_j > 0$ or $\beta_i - \beta_j > 0$ from the subgraph of $F(\beta, \sigma)$ at vertex set $\tilde{\xi}$. By construction and the definition of $B_n$, $\Pi(\xi^*)$ is nested in $\Pi^*(\xi^*)$ and $\tilde{K}_{\tilde{\xi}} \leq (1 + c_\delta)g_n$.

For any $(\beta, \sigma) \in C_n^{(1)}$, we can show that

$$
\Pr \left\{ \|\beta_{\xi} - \beta_{\tilde{\xi}}\| \leq \sigma^* \varepsilon_n \mid \beta, \sigma^2 \right\} 
$$

$$
= \Pr \left\{ \|\tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}}\|^{-1} \tilde{X}_{\tilde{\xi}}^T \varepsilon_z + \tilde{\beta}_{\xi} + (\tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}})^{-1} \tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} - \tilde{\beta}_{\tilde{\xi}} \| \leq \sigma^* \varepsilon_n \right\}
$$

$$
\leq \Pr \left\{ \|\tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}}\|^{-1} \tilde{X}_{\tilde{\xi}}^T \varepsilon_z \| \geq \|\tilde{\beta}_{\xi} - \beta_{\tilde{\xi}}\| - \|((\tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}})^{-1} \tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} - \sigma^* \varepsilon_n) \right\}/\sigma
$$

$$
\leq \Pr \left\{ \|\tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}}\|^{-1} \tilde{X}_{\tilde{\xi}}^T \varepsilon_z \| \geq M_2 \varepsilon_n \right\}
$$

$$
\leq \exp(-c'_{22}n\varepsilon_n^2)
$$

(A11)

(A12)

where $z := \varepsilon/\sigma$ following $N(0, 1)$. The last inequality in (A12) is from Lemma 1. The second last inequality in (A11) holds because $\|\tilde{\beta}_{\xi} - \beta_{\tilde{\xi}}\| \geq \|\beta_{\xi} - \beta_{\tilde{\xi}}\| - \|\beta_{\tilde{\xi}}\| \geq M_1 \sigma^* \varepsilon_n - p_n(\sigma \varepsilon_n/p) \geq M_1 \sigma^* \varepsilon_n - \sigma^* \varepsilon_n \sqrt{(1 + \varepsilon_n)/(1 - \varepsilon_n)}$, and $\|((\tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}})^{-1} \tilde{X}_{\tilde{\xi}}^T \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} - \sigma^* \varepsilon_n) \leq \sqrt{1/(n\lambda_0)} \sqrt{\varepsilon_n} \sqrt{\sigma} \varepsilon_n \leq c'_{23} \sigma^* \varepsilon_n$ by Assumption (A-2).

For any $(\beta, \sigma) \in C_n^{(2)}$, from linear regression results, we have $\left\| y - \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} \right\|^2 \sim \sigma_2^2 \chi_n^2 - K_{\tilde{\xi}}(\kappa)$, where the noncentral parameter $\kappa = \left\| \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} / \sigma \right\|^2 \leq c_n \varepsilon_n^2$. Therefore,

$$
\Pr_{(\beta, \sigma)} \left( \left| \frac{\sigma_2^2(y) - \sigma^2}{\sigma^2(n - K_{\tilde{\xi}})} \right| < \varepsilon_n \right)
$$

$$
= \Pr_{(\beta, \sigma)} \left( \left| \frac{\left\| y - \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} \right\|^2}{\sigma^2(n - K_{\tilde{\xi}})} - 1 \right| < \varepsilon_n \right)
$$

$$
\leq \Pr_{(\beta, \sigma)} \left( \left| \frac{\left\| y - \tilde{X}_{\tilde{\xi}} \tilde{\beta}_{\tilde{\xi}} \right\|^2}{\sigma^2} - (n - K_{\tilde{\xi}}) \right| > (n - K_{\tilde{\xi}}) \varepsilon_n \right)
$$

$$
\leq \Pr \left( \left| \chi_n^2 - K_{\tilde{\xi}}(\kappa) - (n - K_{\tilde{\xi}}) \right| > (n - K_{\tilde{\xi}}) \varepsilon_n \right)
$$

$$
\leq \exp \left( -c'_{24} n \varepsilon_n^2 \right),
$$

(A13)

for some constant $c'_{24} > 0$ and large $n$.

Therefore,

$$
\sup_{(\beta, \sigma) \in C_n} E_{(\beta, \sigma)} \{1 - \phi(y)\} \leq \max \left\{ \sup_{(\beta, \sigma) \in C_n^{(1)}} E_{(\beta, \sigma)} \{1 - \phi(y)\}, \sup_{(\beta, \sigma) \in C_n^{(2)}} E_{(\beta, \sigma)} \{1 - \phi(y)\} \right\}
$$

$$
\leq \max \left\{ \exp \left( -c'_{22} n \varepsilon_n^2 \right), \exp \left( -c'_{24} n \varepsilon_n^2 \right) \right\}
$$

$$
\leq \exp \left( -c_{22} n \varepsilon_n^2 \right),
$$

(A14)

Part 3 (Evidence lower bound): Recall $m(y)$ is the marginal likelihood and $f^*(y)$ is the true likelihood. We let $e^* = y - X\beta^*$ be the vector of error terms. Under our model,

$$
\frac{m(y)}{f^*(y)} = \int \exp(R_n) \pi(\beta, \sigma^2) d\beta d\sigma^2,
$$

where $R_n$ is the log likelihood ratio $-\frac{1}{2\sigma^2} \left\| X\beta^* + e^* - X\beta \right\|^2 + \frac{\left\| e^* \right\|^2}{2\sigma^2} - n \log \frac{1}{\sigma}$.

In Part 3, we seek to prove

$$
\Pr \left\{ \frac{m(y)}{f^*(y)} \geq \exp(-c_{31} n \varepsilon_n^2) \right\} \geq 1 - \exp(-c_{32} n \varepsilon_n^2),
$$

(A15)
for some constants $c_{31}$ and $c_{32}$.

We will first show that (A15) holds under the event $E_1 = \{ \| \epsilon^* / \sigma^* \|^2 \leq (1 + c_{31})n \} \cap \{ \epsilon^T X \|_{\infty} \leq c_{31} n \epsilon_n \}$ and then show that $E_1$ holds with a large probability.

To prove (A15), rewrite

$$R_n = \frac{\| X \beta^* - X \beta \|_i^2}{2 \sigma^2} - \frac{(X \beta^* - X \beta)^T \epsilon^*}{\sigma^2} + \| \epsilon^* \|^2 \left( \frac{1}{2 \sigma^2} - \frac{1}{2 \sigma^2} \right) - \frac{n}{2} \log \frac{\sigma^2}{\sigma^2} \right)_{>0} > -\delta_{31} n \epsilon_n^2 / 2 \quad (A16)$$

Under the event of $E_1$, when $H_2 := \{ \sigma^2 \in [\sigma_{\epsilon^2}, \sigma_{\epsilon^2}(1 + \delta_{31}r_2^2)] \} \cap \{ (\beta^* - \beta)/\sigma \|_1 \leq 2 \delta_{32} \}$ holds for some constants $\delta_{31}$ and $\delta_{32}$, we can show that $I \geq O(-n \epsilon_n^2)$ and $\| I \| \geq O(n \epsilon_n^2)$ using inequality between $l_1$ and $l_2$ norms and Hölder inequality. Thus, $H_1 := \{ R_n \geq -c_{32} n \epsilon_n^2 \}$ is a super-set of $H_2$ for some constant $c_{32}$.

Let $\pi(\sigma^2)$ denote the prior density function for $\sigma^2$. By Assumption (A-4), for some constant $c_\sigma > 0$,

$$\Pr \left\{ \sigma^2 \in [\sigma_{\epsilon^2}, \sigma_{\epsilon^2}(1 + \delta_{31}r_2^2)] \right\} \geq \delta_{31} \sigma_{\epsilon^2} \sigma_{[\sigma_{\epsilon^2}, \sigma_{\epsilon^2}(1 + \delta_{31}r_2^2)]} \pi(\sigma^2) \geq \exp(-c_\sigma \epsilon_n^2) \quad (A17)$$

Let $\Pi$ denote an arbitrary partition of $\{1, \ldots, p\}$ whose corresponding partition of $\xi^*$ is determined by removing those edges with $\beta_1^* - \beta_2^* > 0$ from the subgraph of any $F$ compatible with $\Pi$ at vertex set $\xi^*$.

Also let $\tilde{K}, \tilde{\beta}$ and $\tilde{\beta}^*$ denote its number of clusters, the transformed coefficients and true coefficients, respectively. We now analyze the prior of $\beta$. Without loss of generality, assume the first $\tilde{k}$ clusters of $\Pi$ have nonzero $\beta^*$. We write $\| (\beta^* - \beta)/\sigma \|_1 = \sum_{j=1}^{k-1} \sqrt{|\tilde{C}_j|} |\tilde{\beta}_j - \tilde{\beta}_j^*|/\sigma^2 + \sum_{j=\tilde{k}+1}^{K} \sqrt{|\tilde{C}_j|} |\tilde{\beta}_j|/\sigma$. Therefore,

$$\| (\beta^* - \beta)/\sigma \|_1 \geq 2 \delta_{32} \epsilon_n \left\{ \begin{array}{l} \tilde{\beta}_j/\sigma \leq \delta_{32} \epsilon_n / \sqrt{\tilde{K}p} \quad \text{for all } j = 1, \ldots, \tilde{k} \end{array} \right\} \cap \left\{ \begin{array}{l} \tilde{\beta}_j/\sigma \geq \delta_{32} \epsilon_n / \sqrt{\tilde{k}1|\xi^*|} \quad \text{for all } j = 1, \ldots, \tilde{k} \end{array} \right\}$$

from the inequality between $l_1$ and $l_2$ norms.

Note that $\tilde{k} \leq g_n^*$ and $\tilde{K} \leq p$. Conditional on $\sigma^2 \in [\sigma_{\epsilon^2}, \sigma_{\epsilon^2}(1 + \delta_{31}r_2^2)]$ and $\Pi$, we have

$$\Pr \left\{ \| \tilde{\beta}_j/\sigma \| \leq \delta_{32} \epsilon_n / \sqrt{\tilde{K}p} \quad \text{for all } j = 1, \ldots, \tilde{K} \right\} \geq (1 - p^{1-c_{\epsilon^2}} \sqrt{\tilde{K}p}) \rightarrow 1 \quad (A18)$$

and

$$\Pr \left\{ \| \tilde{\beta}_j/\sigma \| \geq \delta_{32} \epsilon_n / \sqrt{\tilde{k}1|\xi^*|} \quad \text{for all } j = 1, \ldots, \tilde{k} \right\} \geq \exp(-c_\sigma \epsilon_n^2) \quad (A19)$$

from Lemma 2 and Assumption (A-1). Combining (A17), (A18) and (A19), we can show that $\Pr(H_2-\Pi) \geq \exp(-c_{33} n \epsilon_n^2)$.

Also notice Pr($\Pi | F$) $= \sum_{K \geq k} n_c - n_c^*$ Pr($k + 1, n_c - n_c^*$) Pr($\beta^* = \tilde{K}) (p - n_c - k^1 + n_c^1) (p - n_c) (K - n_c) - 1$, where $n_c^*$ is the number of components in the original graph on $\xi^*$. By Assumption (A-1),

$$\log \Pr(k = \tilde{k}^1 + n_c - n_c^*) \geq \log \left( \frac{(1 - c)^{2g_n^*} + \sum_{k=1}^{p_n}(1 - c)^k} {2g_n^* - 1} \log (1 - c) + 2c \log (1 - c)^p \right) \geq -2c_a g_n^* \log (1 - c) \quad (A20)$$
In addition,
\[ -\log \left( \frac{p - n_c}{k_1 - n_c^*} \right) \geq -g_n^* \log p. \] (A21)

Therefore, \( \Pr(H_2|F) \geq \Pr(H_2|\tilde{\Pi}) \Pr(\tilde{\Pi}|F) \geq \exp \left( -c'_{34} n \varepsilon_n^2 \right). \) Since \( c'_{34} \) does not depend on the choice of \( F \), we have \( \Pr(H_2) \geq \exp \left( -c'_{34} n \varepsilon_n^2 \right). \)

Recall under \( E_1 \), \( H_2 \) holds and thus \( R_n \geq c'_{32} n \varepsilon_n^2 \) since \( H_1 \supset H_2 \). It follows that,
\[
\frac{m(y)}{f^*(y)} \geq \Pr(H_2) \exp(R_n) \geq \exp(-c'_{32} n \varepsilon_n^2) \Pr(H_2) \geq \exp(-c_{31} n \varepsilon_n^2). \] (A22)

Finally, from Lemma 1 and the tail bound for maximum of sub-Gaussian random variables, we can prove that \( \Pr(E_1) \geq 1 - \exp \{ -c_{32} n \varepsilon_n^2 \} \). Combining with (A22), we proved (A15).

**Combining three parts:** Finally, we use the results of (A9) from Part 1, (A10) and (A14) from Part 2, (A15) from Part 3 and apply Lemma 3 to obtain
\[
E^* \left\{ \pi_n \left( \| \beta - \beta^* \|_2 \geq M_1 \sigma^* \varepsilon_n \mid y \right) \right\} \leq E^* \left\{ \pi_n \left( C_n \cup B_n \mid y \right) \right\} \leq \exp(-c_e n \varepsilon_n^2)
\]
for some constant \( c_e > 0 \). The result in Theorem 1 then follows from Markov inequality and Borel-Cantelli lemma, which completes the proof of Theorem 1.

\[ \square \]
A5 Hyperparameter Selection and Computational Efficiency Analysis

A5.1 Hyperparameter Sensitivity Analysis

First we present additional simulation study results in Section 4.1 with different sets of hyperparameters. For T-LoHo, 6 choices of \((\tau_0, c) \in \{0.1, 1\} \times \{0.1, 0.5, 0.9\}\) are considered. Note that we used posterior median estimate for \(\hat{\beta}^{\text{T-LoHo}}\), because T-LoHo gives multimodal posterior distribution near the cluster boundary where using posterior mean can lead to misleading summary. For soft-thresholded GP (STGP) (Kang et al., 2018), we used \(15 \times 15\) equally spaced grid of knots with bandwidth set to the minimum distance between knots and same prior specification of section 4.1 therein. We used posterior mean to get \(\hat{\beta}^{\text{STGP}}\) where \(\hat{\beta}^{\text{STGP}} = 0\) if posterior probability of a nonzero \(\hat{\beta}^{\text{STGP}}\) is less than 1/2. We ran total 50,000 MCMC iterations with 10,000 burn-in and 10 thin-in rate to collect 4,000 posterior samples for both T-LoHo and STGP. For sparse Fused lasso (FL) (Tibshirani et al., 2011), we used path algorithm (Arnold and Tibshirani, 2020) to get FL estimate

\[
\hat{\beta}_{FL} = \arg\min_{\beta} \left\{ 0.5\|y - X\beta\|_2^2 + \lambda_{FL} \sum_{(j,l) \in E} |\beta_j - \beta_l| + \gamma_{FL} \cdot \lambda_{FL} \sum_{j=1}^p |\beta_j| \right\}
\]

where we considered 3 candidates of \(\gamma_{FL} \in \{0.2, 1, 5\}\) and selected \(\lambda_{FL}\) among 2000 number of steps using Bayesian information criteria. For graph octagonal shrinkage and clustering algorithm for regression (GOSCAR) (Yang et al., 2012), \(\lambda_1 (\ell_1\) regularizator parameter) and \(\lambda_2 (\ell_\infty\) regularizator parameter) are both set as 0.8 \(\max\{|\beta_i|/|E|\}\) from the true value of \(\max\{|\beta_i|\}\). For Bayesian graph Laplacian (Liu et al., 2014), we used EM algorithm of Chakraborty and Lozano (2019) with same prior specification of Section 2.6 therein.

We report mean squared prediction error (MSPE) \(\frac{1}{n_{\text{test}}} \| \tilde{y}_{\text{test}} - \hat{y}_{\text{test}} \|_2^2\) based on a test set with size \(n_{\text{test}} = 1000\) and Rand index (RI) (Rand, 1971) which measures the clustering accuracy. Table A1 shows comparison results with different levels of dependencies and SNR. Based on 100 replicated simulations, average MSPE and RI values are reported under 6 different settings of \((\tau_0, c)\) for T-LoHo, one STGP, and 3 different settings of \(\gamma\) for FL. In Section 4.1, we report T-LoHo result with \((\tau_0, c) = (1, 0.5)\) and FL result with \(\gamma_{FL} = 0.2\).

Table A1: Performance comparison based on average MSPE and RI over 100 replicated simulations. Standard error is given in parentheses. RI = 1 indicates exact recovery of the true cluster.

<table>
<thead>
<tr>
<th>(\vartheta), SNR</th>
<th>T-LoHo with different ((\tau_0, c)) settings</th>
<th>STGP</th>
<th>FL with different (\gamma_{FL}) settings</th>
<th>GOSCAR</th>
<th>BGL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.1, 0.1)</td>
<td>(0.1, 0.5)</td>
<td>(0.1, 0.9)</td>
<td>(1, 0.1)</td>
<td>(1, 0.5)</td>
</tr>
<tr>
<td>MSPE</td>
<td></td>
<td>57.1(2.6)</td>
<td>75.1(3.9)</td>
<td>101.2(4.4)</td>
<td><strong>52.1(2.2)</strong></td>
</tr>
<tr>
<td>RI</td>
<td></td>
<td>0.43(0.007)</td>
<td>0.85(0.009)</td>
<td>0.81(0.009)</td>
<td>0.58(0.018)</td>
</tr>
<tr>
<td></td>
<td>(0.1, 0.5)</td>
<td>27.6(1.4)</td>
<td>30.8(1.6)</td>
<td>46.5(4.1)</td>
<td>27.1(1.1)</td>
</tr>
<tr>
<td>MSPE</td>
<td></td>
<td>241.3(2.2)</td>
<td>115.7(3.6)</td>
<td>128.5(4.2)</td>
<td><strong>221.9(7.3)</strong></td>
</tr>
<tr>
<td>RI</td>
<td></td>
<td>0.72(0.009)</td>
<td>0.47(0.012)</td>
<td>0.66(0.004)</td>
<td>0.66(0.002)</td>
</tr>
<tr>
<td></td>
<td>(0.1, 0.9)</td>
<td>3, 4 0.72(0.010)</td>
<td>0.46(0.007)</td>
<td>0.69(0.004)</td>
<td>0.66(0.002)</td>
</tr>
<tr>
<td></td>
<td>(1, 0.1)</td>
<td>3, 4 0.79(0.004)</td>
<td>0.58(0.012)</td>
<td>0.78(0.007)</td>
<td>0.65(0.006)</td>
</tr>
<tr>
<td></td>
<td>(1, 0.5)</td>
<td>3, 4 0.80(0.003)</td>
<td>0.57(0.011)</td>
<td>0.82(0.007)</td>
<td>0.85(0.003)</td>
</tr>
<tr>
<td></td>
<td>(1, 0.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(\tau_0, c) = (0.1, 0.1)

\begin{figure}[h]
\centering
\begin{subfigure}{0.4\textwidth}
\centering
(\tau_0, c) = (0.1, 0.1)
\end{subfigure} \hfill
\begin{subfigure}{0.4\textwidth}
\centering
(\tau_0, c) = (1, 0.1)
\end{subfigure}
\begin{subfigure}{0.4\textwidth}
\centering
(\tau_0, c) = (0.1, 0.5)
\end{subfigure} \hfill
\begin{subfigure}{0.4\textwidth}
\centering
(\tau_0, c) = (1, 0.5)
\end{subfigure}
\caption{T-LoHo posterior median estimate of $\beta$ when ($\tilde{\beta}$,SNR) = (0.4) with different choice of hyperparameters $(\tau_0, c) \in \{(0.1, 0.1), (1, 0.1), (0.1, 0.5), (1, 0.5)\}$. Black boundaries illustrates the point estimate of cluster obtained from MCMC samples, based on Dahl’s method Dahl (2006).}
\end{figure}

Figure A2: Posterior distribution of $\beta_j$ at two different locations, (a) and (b) as displayed in bottom panel of fig.A1. Red line is when $\tau_0 = 0.1$, blue is when $\tau_0 = 1$. Dashed line denotes the true value.
Selection and sensitivity analysis of $\tau_0$  
Hyperparameter $\tau_0 > 0$ controls the scale of the global shrinkage parameter, $\tau \sim C^+((0, \tau_0))$. As $\tau_0$ approaches to 0, posterior distribution of $\beta$ more concentrates at 0 for zero coefficients as we can see in the left panel of fig. A2. However at the same time, a too small $\tau_0$ can introduce a stronger bias towards zero for nonzero signals as shown in the right panel of fig. A2 (although this bias may disapper as the magnitude of signal increases). Thus, there is a trade-off between shrinkage and bias when selecting $\tau_0$. In practice, if $\beta$ is assumed to be very sparse (i.e. contains many zeros) then a small $\tau_0$ is preferred because the benefit of shrinking surpasses the loss due to bias. Under the formulation of T-LoHo, it is desirable to use a small $\tau_0$ if the clusterwise parameter $\beta$ is assumed to be very sparse. Recently, Piironen and Vehtari (2017) proposed to select $\tau_0$ based on the prior guess for the number of relevant variables. Following their approach, we suggest $\tau_0 = p_0/(K_0 - p_0)$ where $p_0$ is a prior guess for the number of nonzero clusters, $K_0$ is a prior guess for the total number of clusters. In our simulation setting, $\tau_0^* = 2/(3 - 2) = 2$ would be an appropriate choice of $\tau_0$ if true is known.

According to table A1, when we have an independent design ($\vartheta = 0$), $\tau_0 = 1$ gives slightly better results in terms of MSPE compared to $\tau_0 = 0.1$. This is because the shrinkage effect is not well reflected in the posterior median estimate $\hat{\beta}$, while the bias is more emphasized when $\tau_0$ gets smaller. However when we have a dependent design ($\vartheta = 3$), $\tau_0 = 0.1$ gives slightly better results in terms of MSPE compared to $\tau_0 = 1$. This may be explained by the fact that dependency of $X$ leads to many falsely identified clusters, where the benefit of shrinking those is more emphasized.

Selection and sensitivity analysis of $c$  
Hyperparameter $c$ controls the geometric decay rate in the prior of $K$, i.e., $\Pr(K=k) \propto (1-c)^k$. As $c$ becomes closer to 1, it more penalizes on the cluster generation and gives fewer clusters. Assumption (A-4) of Theorem 1 requires $1-c \geq p^{-\alpha}$, which suggests that one should not choose the value of $c$ too close to 1 unless the number of vertices $p$ is very large. Indeed, we notice that when $c = 0.9$, the MSPE increases according to table A1.

Table A1 also shows that Rand index sharply drops down when $c = 0.1$, and it is even more emphasized when $\tau_0 = 0.1$. Indeed, at the top panel of fig. A1 we can see that the cluster estimate is highly compartmentalized. This phenomenon can be characterized with our previous discussion in Section 3.2, where we describe clustering as a trade-off between the Bayes factor $\mathcal{L} = p(y|M_2)/p(y|M_1)$ and the the penalization prior $p(K) \propto (1-c)^K$. As $\tau_0$ becomes close to 0, induced density $\pi_\Delta$ has a sharper spike around the origin, which makes the alternative $\delta \sim \pi_\Delta$ behave more similar to the null $\delta = 0$. As $c$ becomes too close to 0, the effect of penalization prior $p(K)$ fades out, which leads to many redundant cluster estimates. Therefore, we recommend not to choose $c$ too close to 0.

In summary, we suggest to use $c = 0.5$ as a default and not to use too small or too large $c$. If the number of vertices in $G$ is sufficiently large. Assumption (A-4) gives a more wider choice of $c$, and users may choose a different value of $c$ based on their prior belief on the model size or some model selection criteria such as the Watanabe-Akaike information criterion (WAIC; Watanabe, 2010).

A5.2 Simulation Studies With More Clusters

We performed additional simulation studies when there are multiple clusters in the data. Two scenarios are considered: 1) four nonzero clusters with values $-2, -1, 1, 2$ respectively; 2) eight nonzero clusters with values $-3, -3, -2, -1, 1, 2, 3, 3$ respectively. See fig. A3 for the graphical illustration. Similar as before, we compared the MSPE and Rand index under the setting of $n = 250$ and $p = 900$ in tables A2 and A3. The T-LoHo hyperparameters are set as $(\tau_0, c) = (1, 0.5)$ and $\gamma_{FL} \in \{0.2, 1.5\}$. When sample size $n$ becomes smaller, it will be more challenging to identify clusters when the number of clusters is larger. In theorem 1, $g_n^*$, the size of the spanning tree space compatible with nonzero clusters, increases as the number of nonzero clusters increases. Assumption (A-1) states that it should be upper bounded by $n/\log p$ asymptotically, which may not hold when $n$ is small. Indeed, we observed that T-LoHo, FL, GOSCAR, and BGL all suffer from a much poorer performance.
Table A2: Performance comparison based on average MSPE and RI over 100 replicated simulations. Standard error is given in parentheses and time is in sec. RI = 1 indicates exact recovery of true cluster.

<table>
<thead>
<tr>
<th>K = 4</th>
<th>SNR</th>
<th>T-LoHo</th>
<th>STGP</th>
<th>FL (0.2)</th>
<th>FL (1)</th>
<th>FL (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>85.2(3.1)</td>
<td>190.7(1.9)</td>
<td>167.3(4.4)</td>
<td>332.6(8.4)</td>
<td>485.2(4.7)</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>34.9(1.4)</td>
<td>172.3(1.4)</td>
<td>91.7(2.8)</td>
<td>197.7(7.2)</td>
<td>365.0(3.8)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>399.0(14.0)</td>
<td>441.8(8.4)</td>
<td>566.7(21.8)</td>
<td>656.4(23.9)</td>
<td>1050.2(32.4)</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>206.5(7.1)</td>
<td>361.0(7.5)</td>
<td>304.6(9.4)</td>
<td>377.9(12.5)</td>
<td>649.9(18.7)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RI</th>
<th>0</th>
<th>0.91(0.007)</th>
<th>0.65(0.010)</th>
<th>0.83(0.008)</th>
<th>0.76(0.008)</th>
<th>0.58(0.002)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>0.95(0.005)</td>
<td>0.63(0.008)</td>
<td>0.86(0.009)</td>
<td>0.77(0.013)</td>
<td>0.61(0.002)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.88(0.004)</td>
<td>0.77(0.003)</td>
<td>0.61(0.011)</td>
<td>0.79(0.006)</td>
<td>0.78(0.003)</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.92(0.003)</td>
<td>0.75(0.004)</td>
<td>0.60(0.011)</td>
<td>0.80(0.006)</td>
<td>0.82(0.002)</td>
</tr>
</tbody>
</table>

Table A3: Performance comparison based on average MSPE and RI over 100 replicated simulations. Standard error is given in parentheses and time is in sec. RI = 1 indicates exact recovery of true cluster.

<table>
<thead>
<tr>
<th>K = 8</th>
<th>SNR</th>
<th>T-LoHo</th>
<th>STGP</th>
<th>FL (0.2)</th>
<th>FL (1)</th>
<th>FL (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>427.2(9.7)</td>
<td>414.5(4.3)</td>
<td>658.0(20.1)</td>
<td>1282.1(23.3)</td>
<td>1574.7(13.9)</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>251.8(6.2)</td>
<td>352.7(3.3)</td>
<td>405.0(12.4)</td>
<td>934.4(33.0)</td>
<td>1271.0(11.2)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2285.1(61.7)</td>
<td>1424.9(28.8)</td>
<td>2258.7(71.7)</td>
<td>2814.4(81.4)</td>
<td>4747.6(146.3)</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1189.3(31.4)</td>
<td>978.1(20.1)</td>
<td>1364.8(47.2)</td>
<td>1723.1(57.1)</td>
<td>2944.4(87.3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RI</th>
<th>0</th>
<th>0.76(0.006)</th>
<th>0.59(0.006)</th>
<th>0.72(0.006)</th>
<th>0.54(0.008)</th>
<th>0.57(0.004)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>0.84(0.006)</td>
<td>0.61(0.005)</td>
<td>0.78(0.005)</td>
<td>0.64(0.009)</td>
<td>0.61(0.002)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.71(0.006)</td>
<td>0.62(0.005)</td>
<td>0.67(0.006)</td>
<td>0.74(0.003)</td>
<td>0.67(0.004)</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.78(0.006)</td>
<td>0.64(0.004)</td>
<td>0.68(0.005)</td>
<td>0.77(0.004)</td>
<td>0.73(0.003)</td>
</tr>
</tbody>
</table>

Figure A3: True and fitted results with four (first row), eight (second row) nonzero clusters. Here (θ, SNR) = (0, 4). (a) True coefficient image β; (b) T-LoHo estimate with (τ0, c) = (1, 0.5); (c) STGP estimate after thresholding; (d) FL estimate with γFL = 0.2.

The performance of all methods deteriorates as the number of nonzero clusters increases. The results are not surprising because, for many high-dimensional regularization methods, it is known that the learning rate often depends on the size of the true model (the number of clusters in our case) relative to the sample size. Except in the case when K = 8 with θ = 3, T-LoHo still outperforms other baselines in terms of both MSPE and Rand index. It will be of interest to improve the performance of T-LoHo when predictors are correlated, sparsity level is relatively low, and it has many nonzero clusters. We left these as a future working direction.
References


