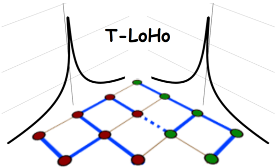
T-LoHo: A Bayesian Regularization Model for Structured Sparsity and Smoothness on Graphs

Changwoo Lee, Zhao Tang Luo and Huiyan Sang

Department of Statistics
Texas A&M University



NeruIPS 2021

Beyond Sparsity in High-dimensional Models

High-dimensional models: the number of parameters p exceeds the sample size n

- **Sparsity Assumption**
  - Assumes p-dimensional parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ has many zero components
  - Avoids overfitting & Improves predictive accuracy and interpretation
- **Sparse Homogeneity Assumption**
  - Assumes $\boldsymbol{\beta}$ has clustered patterns and many possibly clustered zeros
  - Plausible when $\boldsymbol{\beta}$ has a pre-known structure (e.g. time, image, . . .)

Sparsity Assumption:
'A needle in a haystack'

Sparse Homogeneity Assumption:
'Bunches of needles in a haystack'

Sparse Homogeneity and Graph Structured Parameters

Examples of models under the sparse homogeneity assumption:

- **Fused lasso(FL)**[Tibshirani et al., 2005]: for time-neighboring coefficients $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ 0.5\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \textstyle\sum_{j=2}^{p} |\beta_j - \beta_{j-1}| + \gamma \textstyle\sum_{j=1}^{p} |\beta_j| \right\}$$

- **Generalized fused lasso**[Tibshirani et al., 2011]: consider an undirected graph $G = (V, E)$ with $|V| = p$ which represents a pre-known structure of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ 0.5\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{(j,k)\in E} |\beta_j - \beta_k| + \gamma \sum_{j=1}^{p} |\beta_j| \right\}$$

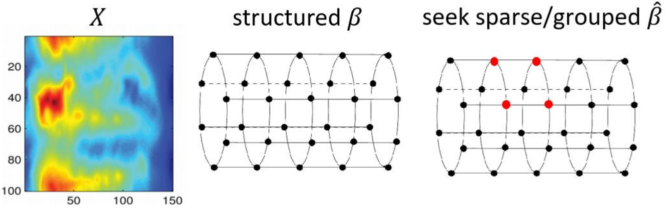


Figure 6. Observed left hippocampus images. [Wang et al., 2017]

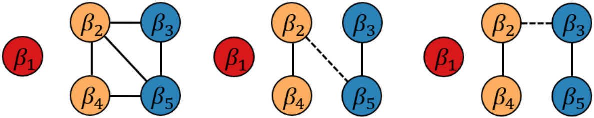
Bayesian Model under Sparse Homogeneity

Here we propose a prior model on β which induces clustered sparsity. It delivers full Bayesian inference for model parameters, including the number of clusters.

- Let $\Pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be a graph partition of G.
- First we introduce a tree-based prior on Π using random spanning tree/forest. Here Π can be represented through **cuts of spanning forest** of G:

Proposition 1(full coverage of graph partition with cuts of spanning forest)

Let $G = (V, E)$ be a graph with n_c connected components and $\Pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be an arbitrary graph partition of G. There exists a spanning forest $\mathcal{F} = (V, E^F)$ with $|E^F| = |V| - n_c$, and a set of cut-edges $E^C \subset E^F$ with $|E^C| = K - n_c$ such that the induced cut of $\mathcal{F}$ is Π.



Graph G and partition $\Pi = \{C_1, C_2, C_3\}$ | An example $\mathcal{F}$ and cut-edge(dashed) | Another example of $\mathcal{F}$ and cut-edge(dashed)

A Bayesian Random Spanning Forest Partition Prior

We specify a prior on Π through spanning forest $\mathcal{F}$ and the number of clusters K. Specifically, we have a prior on spanning forest space by choosing $\mathcal{F}$ be the minimum spanning forest of G with random edge weights $w_{ij} \overset{iid}{\sim} \text{Unif}(0,1)$:

$$
\begin{aligned}
\mathcal{F} &= \text{MSF}(\{w_{ij}\}), \quad w_{ij} \overset{iid}{\sim} \text{Unif}(0,1) && \text{(uniform edge weights on a graph)}\\
p(K=k) &\propto (1-c)^k, \quad k = n_c, \ldots, K && \text{(geometric prior on \#(clusters))}\\
p(\Pi \mid \mathcal{F}, K) &\propto 1(|\Pi| = K \text{ and is induced by } \mathcal{F}) && \text{(uniformly select } K - n_c \text{ cut-edges)}
\end{aligned}
$$

- Extension of [Luo et al., 2021], but different from [Teixeira et al., 2019].
- $c \in [0,1)$ controls the model size, c closer to 1 penalizes large #(clusters).
- After $\mathcal{F}$ and K are given, Π is determined by selecting $K - n_c$ cut-edges uniformly at random to get a graph partition Π of size K.

Bayesian Graph Structured Sparsity

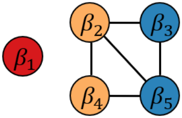
Given a graph partition Π, we can construct a $K \times p$ matrix Φ from Π:

$$\Phi_{kj} = 1/\sqrt{|\mathcal{C}_k|} \text{ if } j \in \mathcal{C}_k \text{ and } 0 \text{ otherwise}, \quad k = 1, \ldots, K, \; j = 1, \ldots, p$$

We propose to use horseshoe prior[Carvalho et al., 2010] to induce sparsity:

$$\boldsymbol{\beta} \,|\, \sigma^2, \tau^2, \Lambda, \Pi \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\tau^2 \underbrace{\Phi^\top \Lambda \Phi}_{\text{rank } K}), \; \Lambda := \mathrm{diag}(\lambda_1^2, \ldots, \lambda_K^2)$$

$$\lambda_k \overset{iid}{\sim} C^+(0,1), \quad \tau \sim C^+(0, \tau_0), \quad p(\sigma^2) \propto 1/\sigma^2$$



Graph G and partition $\Pi = \{C_1, C_2, C_3\}$

$$\Phi = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & 0 & 1/\sqrt{2} & 0 \\ 0 & 0 & 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}_{3\times 5}$$

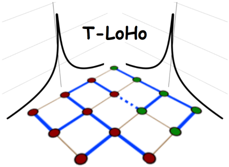
$$\Phi^\top \Lambda \Phi = \begin{bmatrix} \lambda_1^2 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2^2/2 & 0 & \lambda_2^2/2 & 0 \\ 0 & 0 & \lambda_3^2/2 & 0 & \lambda_3^2/2 \\ 0 & \lambda_2^2/2 & 0 & \lambda_2^2/2 & 0 \\ 0 & 0 & \lambda_3^2/2 & 0 & \lambda_3^2/2 \end{bmatrix}_{5\times 5}$$

Figure: Illustrative example of graph partitioning and corresponding parameters when $\boldsymbol{\beta} \in \mathbb{R}^5$ forms $K = 3$ clusters, $\mathcal{C}_1 = \{1\}, \mathcal{C}_2 = \{2, 4\}, \mathcal{C}_3 = \{3, 5\}$.

T-LoHo: Tree-based Low-rank Horseshoe

- The $p \times p$ covariance matrix $\Phi^\top \Lambda \Phi$ has a low-rank(rank $K \ll p$)
- Probability density of $\mathcal{N}_p(\mathbf{0}, \sigma^2\tau^2\Phi^\top\Lambda\Phi)$ lies on rowsp(Φ) with dimension K
- Row space of Φ restricts $\beta_i = \beta_j$ if β_i and β_j lies in a same cluster
- By considering transformation $\tilde{\boldsymbol{\beta}} := \Phi\boldsymbol{\beta}$, we have $\tilde{\boldsymbol{\beta}} \sim \mathcal{N}_K(\mathbf{0}, \sigma^2\tau^2\boldsymbol{\Lambda})$
- Since $\Phi^\top$ is M-P pseudoinverse of Φ, we can recover $\boldsymbol{\beta} = \Phi^\top\tilde{\boldsymbol{\beta}}$

In summary, we propose a Bayesian hierarchical model with (1) Tree-based prior $[\Pi|\mathcal{F}, K][\mathcal{F}][K]$ and (2) Low-rank horseshoe prior $[\beta|\sigma^2, \tau^2, \Lambda, \Pi][\sigma^2][\tau^2][\Lambda]$



T-LoHo: Tree-based Low-rank Horseshoe Model

T-LoHo with linear model

T-LoHo prior can be naturally incorporated into a linear model. With response vector $\boldsymbol{y} \in \mathbb{R}^n$ and column-standardized design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

$$\boldsymbol{\beta} \sim \text{T-LoHo with a (known) undirected graph } G$$

Denote set of parameters $\Theta := (\tilde{\boldsymbol{\beta}}, \sigma^2, \Lambda, \tau, \Pi, K, \mathcal{F})$ and $\tilde{\mathbf{X}} := \mathbf{X}\Phi^\top$ so that $\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \mathbf{X}\Phi^\top\Phi\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$. Then the posterior $p(\Theta|\boldsymbol{y})$ becomes

$$\begin{aligned} p(\Theta|\boldsymbol{y}) \propto \; & \mathcal{N}_n(\boldsymbol{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n) \times \mathcal{N}_K(\tilde{\boldsymbol{\beta}}|\mathbf{0}, \sigma^2\tau^2\Lambda) \times 1/\sigma^2 \\ & \times (1+\tau^2)^{-1} \prod_{k=1}^{K} (1+\lambda_k^2)^{-1} \times \binom{p-n_c}{K-n_c}^{-1} \times (1-c)^K \times 1 \end{aligned}$$

where the last line is the product of priors $p(\tau)\prod_{k=1}^{K} p(\lambda_k) p(\Pi|K, \mathcal{F}) p(K) p(\boldsymbol{W})$.

Posterior Inference

We design a reversible-jump MCMC[Green, 1995] algorithm to sample from $\Theta|\boldsymbol{y}$:

Algorithm 1: One full iteration of RJMCMC posterior sampler

Step 1. Update $\Pi, K, \mathcal{F}$ using collapsed conditional $[\Pi, K, \mathcal{F}|\Lambda, \tau, \boldsymbol{y}]$ where $\tilde{\boldsymbol{\beta}}, \sigma^2$ 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 $(\Pi^\star, K^\star = K+1)$ compatible with $\mathcal{F}$, and accept with probability $\min\{1, \mathcal{A}_a \cdot \mathcal{P}_a \cdot \mathcal{L}_a\}$, where $\mathcal{A}_a$ is prior ratio, $\mathcal{P}_a$ is proposal ratio, $\mathcal{L}_a$ is likelihood ratio.
- 1-b. (*merge*) Propose $(\Pi^\star, K^\star = K-1)$ compatible with $\mathcal{F}$, and accept w.p. $\min\{1, \mathcal{A}_b \cdot \mathcal{P}_b \cdot \mathcal{L}_b\}$.
- 1-c. (*change*) Propose $(\Pi^\star, K^\star = K)$ compatible with $\mathcal{F}$, and accept w.p. $\min\{1, \mathcal{A}_c \cdot \mathcal{P}_c \cdot \mathcal{L}_c\}$.
- 1-d. (*hyper*) Update $\mathcal{F}^\star$ compatible with current Π.

Step 2. Jointly update $(\tau, \sigma^2, \tilde{\boldsymbol{\beta}})$ from $[\tau, \sigma^2, \tilde{\boldsymbol{\beta}} \mid \Lambda, \Pi, K, \mathcal{F}, \boldsymbol{y}]$, by performing:

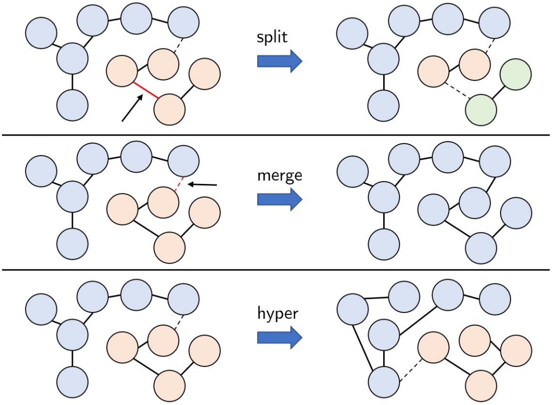
- 2-1. Update τ from $[\tau \mid \Lambda, \Pi, K, \mathcal{F}, \boldsymbol{y}]$ using Metropolis-Hastings sampler,
- 2-2. Update σ^2 from $[\sigma^2 \mid \tau, \Lambda, \Pi, K, \mathcal{F}, \boldsymbol{y}]$ with an inverse gamma distribution,
- 2-3. Update $\tilde{\boldsymbol{\beta}}$ from $[\tilde{\boldsymbol{\beta}} \mid \sigma^2, \tau, \Lambda, \Pi, K, \mathcal{F}, \boldsymbol{y}]$ with a multivariate normal distribution.

Step 3. Update Λ from $[\Lambda \mid \tau, \sigma^2, \tilde{\boldsymbol{\beta}}, \Pi, K, \mathcal{F}, \boldsymbol{y}]$ using slice sampler.

† When $\mathbf{X} = \mathbf{I}_n$(i.e. normal means model), it is possible to integrate out Λ instead of σ^2

Posterior Inference

Step 1 explores graph partitions of G by updating Π, K, and $\mathcal{F}$. It performs one of the four possible moves: (1) split (2) merge (3) change(split & merge) (4) hyper.



Computation Strategies

- In **step 2**: jointly update $\tau, \sigma^2, \boldsymbol{\beta}$ using a blocked Gibbs sampler of [Johndrow et al., 2020] to improve mixing.
- In **step 3**, update $\boldsymbol{\Lambda}$ using a slice sampler.
- **Computation bottleneck:** likelihood calculation which involves calculation of $\boldsymbol{\Sigma}_{n\times n}^{-1}$ and $|\boldsymbol{\Sigma}_{n\times n}|$, where $\boldsymbol{\Sigma}_{n\times n} = \mathbf{I}_n + \tau^2 \tilde{\mathbf{X}}\boldsymbol{\Lambda}\tilde{\mathbf{X}}^\top$ ($O(n^3)$, expensive!).
  - Reduce the rank from n to K by applying Woodbury formula,

$$\boldsymbol{\Sigma}_{n\times n}^{-1} = \mathbf{I}_n - \tilde{\mathbf{X}}(\underbrace{\tau^{-2}\boldsymbol{\Lambda}^{-1} + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}_{(\Sigma^*_{K\times K})})^{-1}\tilde{\mathbf{X}}^\top$$

  - Update the Cholesky decomposition of $\Sigma^*_{k\times k}$ with rank-1 update [Golub and Van Loan, 2013, Sec. 6.5.4] when $\tilde{\mathbf{X}}$ changes. Use the diagonal part updating scheme when τ or $\boldsymbol{\Lambda}$ changes.
  - Computation complexity: $O(\max\{nK, K^3\})$ per iteration, excluding step 1-d which takes $O(m \log p)$ with $m = |E|$.

Clustering effect of T-LoHo

- The difference of parameters $\beta_i - \beta_j$ has an important role in clustering:
  - ℓ_1 penalty(FL)[Tibshirani et al., 2005], ℓ_0 penalty[Fan and Guan, 2018]
  - Prior on $\beta_i - \beta_j$, e.g. Laplace[Kyung et al., 2010], t[Song and Cheng, 2020]
- But putting a prior directly on the difference of parameters $\beta_i - \beta_j$ has many limitations when G has many edges.
- T-LoHo does not directly put prior on the difference $\beta_i - \beta_j$, it puts multivariate prior on $\boldsymbol{\beta}$ with low-rank covariance structure.
- Q. What are the properties of induced prior on $\beta_i - \beta_j$ when $\boldsymbol{\beta} \sim$ T-LoHo?
- A. Horseshoe component of T-LoHo not only introduces shrinkage but also has a clustering effect to facilitate homogeneity, compared to the usual Gaussian prior.

When $\beta_1, \beta_2 \overset{iid}{\sim} \pi_{HS}(\beta) = \int_0^\infty \mathcal{N}(\beta|0, \sigma^2\tau^2\lambda^2) C^+(\lambda|0,1) d\lambda$, it induces prior on standardized difference $\delta = (\beta_1 - \beta_2)/\sigma \sim \pi_\Delta$ where

$$\pi_\Delta(\delta) = \int_0^\infty \mathcal{N}(\delta|0, v) \frac{2}{\pi\tau^2\sqrt{v/\tau^2+1}(v/\tau^2+2)} dv$$

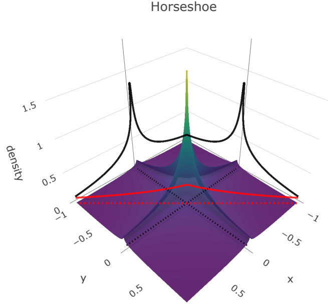
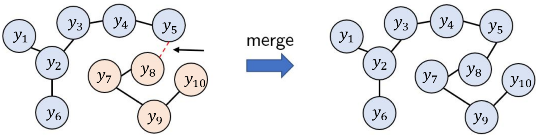


Figure: Joint density $f(x, y) = \pi_{HS}(x)\pi_{HS}(y)$ overlaid with marginal density of $(x - y) \sim \pi_\Delta$ shown as red.

Clustering Effect of T-LoHo

For simplicity, we assume $\mathbf{X} = \mathbf{I}_n$. In step 1-b(merge), we update partition Π based on the acceptance probability $\min\{1, (1-c)^{-1} \times \mathcal{L}\}$. Term $(1-c)$ is from the penalization prior $p(K) \propto (1-c)^K$, and $\mathcal{L}$ is a likelihood ratio. For example,



$$\mathcal{L} = \frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{y}|\mathcal{M}_2)} = \frac{\int p(y_1,\ldots,y_{10}|\beta)p(\beta)d\beta}{\int p(y_1,\ldots,y_6|\beta_1)p(\beta_1)d\beta_1 \int p(y_7,\ldots,y_{10}|\beta_2)p(\beta_2)d\beta_2}$$

$\Longrightarrow$ $\mathcal{L}$ is the Bayes factor of Bayesian two-sample t test[Gönen et al., 2005],

$$\mathcal{M}_1 : \mu_1 = \mu_2, \quad \mathcal{M}_2 : \mu_1 \neq \mu_2, \quad BF_{12} = \frac{P(data\,|\,\mathcal{M}_1)}{P(data\,|\,\mathcal{M}_2)}$$

(Note that for step 1-a(split), $\mathcal{L}$ is simply inverted.)

Clustering Effect of T-LoHo

Following [Gönen et al., 2005], we reparametrize $\delta := (\mu_1 - \mu_2)/\sigma$ with prior $p(\delta)$ which is a parameter of interest and put noninformative prior $p(\frac{\mu_1+\mu_2}{2}, \sigma^2) \propto 1/\sigma^2$ on nuisance parameters. Then the Bayes factor BF_{12} is a function of two-sample t statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p/\sqrt{N}}, \quad \text{where } s_p = \text{pooled sd}, \quad N = (n_1^{-1} + n_2^{-1})^{-1}$$

Q. What are the properties of induced prior on $\beta_i - \beta_j$ when $\boldsymbol{\beta} \sim$ T-LoHo?

We compare BF_{12} when $\delta \sim \mathcal{N}(0,1)$ versus $\delta \sim \pi_\Delta$(induced by T-LoHo).
Scenarios:

- (Balanced groups) $n_1 : n_2 = 1 : 1$ with increasing $\nu = n_1 + n_2 - 2$,
- (Unbalanced groups) $n_1 : n_2 = 9 : 1$ with increasing ν.

Clustering Effect of T-LoHo

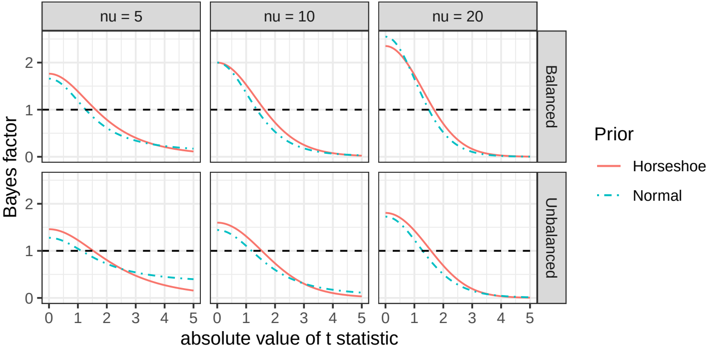


Figure: Comparison of Bayes factor between Horseshoe and Gaussian prior under different (ν, N) settings. Higher Bayes factor implies favoring $H_0 : \mu_1 = \mu_2$

When effect size $|t|$ is small, horseshoe more strongly favors 1-group over 2-groups. In contrast when $|t|$ is big, horseshoe more strongly favors 2-groups over 1-group.

Posterior Consistency

Assumptions:

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

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

(A-3) $\max_j |\tilde{\beta}_j^*|/\sigma^* < L$, where $\log(L) = O(\log p)$.

(A-4) $-\log\tau = O(\log p)$, $\tau < p^{-(2+c_\tau)}\sqrt{g_n^* \log p/n}$, $1 - c \geq p^{-c_\alpha}$, and $\min_{\sigma^2 \in [\sigma^{*2},\, \sigma^{*2}(1+c_\sigma \varepsilon_n^2)]} \pi(\sigma^2) > 0$ for some positive constants c_τ, c_α and c_σ.

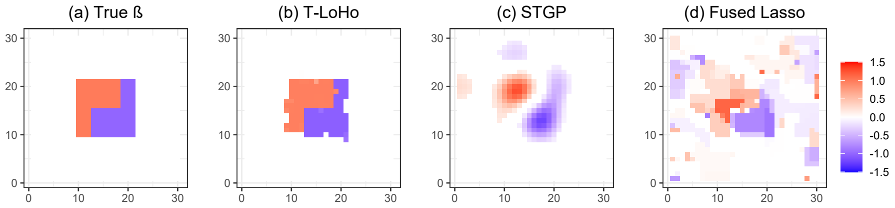
Theorem 1 (Posterior contraction)

Under Assumptions (A-1) to (A-4), there exists a large enough constant $M_1 > 0$ and $\varepsilon_n \asymp \sqrt{g_n^* \log p/n}$ such that the posterior distribution satisfies $\pi_n(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \geq M_1\sigma^*\varepsilon_n \mid \mathbf{y}) \leq \exp(-c_1 n\varepsilon_n^2)$ with probability $1 - \exp(-c_2 n\varepsilon_n^2)$ for some constants $c_1 > 0$ and $c_2 > 0$.

Simulation Settings

- Scalar-on-Image regression model example similar to [Kang et al., 2018]
- Set graph G be a 30×30 lattice graph which represents 2-D image.
- Predictors $\mathbf{X}_i \in \mathbb{R}^{900}$ lying on a graph are generated from iid normal ($\vartheta = 0$) or mean zero Gaussian process(GP) with exponential kernel ($\vartheta > 0$).
- True coefficient $\boldsymbol{\beta} \in \mathbb{R}^{900}$ is sparse(84% zero) with irregular cluster shapes with sharp discontinuities(figure next page)
- Scalar responses $y_i \in \mathbb{R}$, $i = 1, \ldots, 100$ are generated with Gaussian noise with noise variance σ^2 depending on SNR $\in \{2, 4\}$
- Competing methods:
    - Soft-thresholded GP(STGP)[Kang et al., 2018]
    - Sparse fused lasso(FL) [Tibshirani et al., 2011]
    - Graph OSCAR (GOSCAR) [Yang et al., 2012]
    - Bayesian graph Laplacian (BGL) [Liu et al., 2014]
    - Spike-and-slab Laplacian (BayesMSG) [Kim and Gao, 2020]
- Performance measure: Mean square prediction error(MSPE), Rand index(RI)

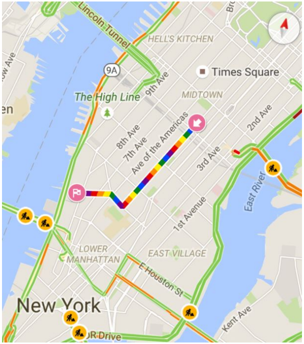
Simulation Results

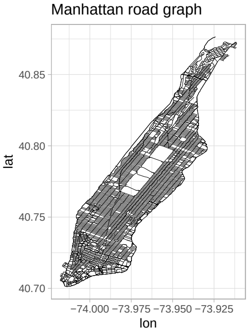


ϑ,SNR	T-LoHo	STGP	FL	GOSCAR	BGL	BayesMSG
MSPE						
0, 2	**68.5(30.0)**	93.4(17.1)	85.0(20.0)	138.2(5.6)	136.2(5.8)	156.1(77.0)
0, 4	**24.4(19.6)**	86.3(15.8)	55.8(14.2)	133.6(5.8)	132.3(5.7)	124.5(27.8)
3, 2	**251.0(112.0)**	278.0(53.0)	341.0(130)	532.3(84.5)	483.2(60.3)	684.5(925.2)
3, 4	**59.7(23.2)**	163.9(21.6)	115.8(36.1)	335.0(48.3)	213.4(27.4)	439.5(74.6)
RI						
0, 2	**0.88(0.06)**	0.72(0.09)	0.47(0.12)	0.28(0.00)	0.28(0.00)	0.42(0.12)
0, 4	**0.95(0.05)**	0.72(0.10)	0.46(0.07)	0.28(0.00)	0.28(0.00)	0.39(0.10)
3, 2	**0.87(0.04)**	0.79(0.04)	0.58(0.12)	0.28(0.00)	0.28(0.00)	0.40(0.13)
3, 4	**0.95(0.02)**	0.80(0.03)	0.57(0.10)	0.28(0.00)	0.28(0.00)	0.29(0.02)
Time						
0, 4	107.9(3.8)	339.9(16.7)	110.4(5.9)	0.11(0.03)	956.2(23.3)	52.9(50.8)

Anomaly Detection in Road Networks

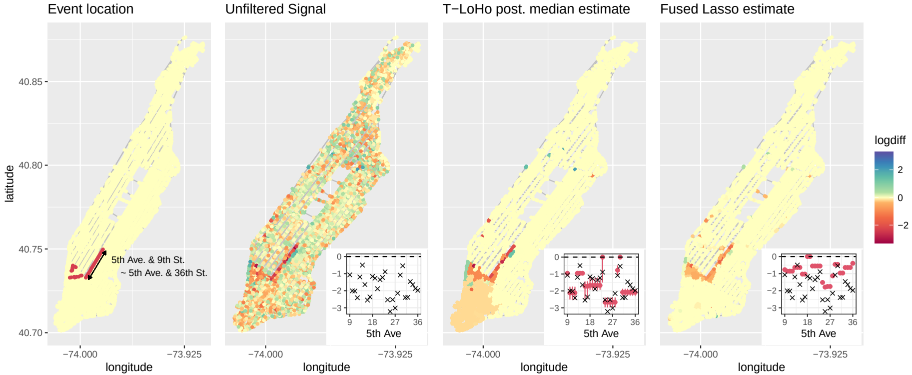
(Revisiting the example of [Wang et al., 2016]) NYC Pride March event held on 12:00 - 14:00, June 26, 2011 which causes traffic congestion. **Goal**: detect clusters on road network which have different taxi pickup/dropoff patterns from usual.





Construct Manhattan road graph $G = (V, E)$ with $|V| = 3748$ and $|E| = 8474$

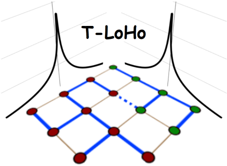
Anomaly Detection in Road Networks



(Left two panels) 2011 NYC pride event route and unfiltered signal. Log-difference value below 0 indicates lower pickup/dropoff frequency than usual. (Right two panels) T-LoHo and FL estimates. (Bottom right subplots) Fitted value comparison zoomed along the parade route, 5th Ave.&9th St. to 5th Ave.&36th St.

Conclusion

- We proposed T-LoHo model, a flexible Bayesian Group Sparsity and Smoothing Regularization method on large graphs.
- Main properties:
  - Can be adapted to various hierarchical model settings;
  - Flexible sparsity and group learning accommodating structural assumptions for easy interpretation;
  - Allows a full Bayesian inference.
- Future work:
  - When we have weighted graph $G = (V, E, w_0)$ instead of $G = (V, E)$.
  - Model variations within active groups.



Thank You!

e-mail: `c.lee@stat.tamu.edu` (Changwoo Lee)

References I

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010).
The horseshoe estimator for sparse signals.
Biometrika, 97(2):465–480.

Fan, Z. and Guan, L. (2018).
Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs.
The Annals of Statistics, 46(6B):3217–3245.

Golub, G. H. and Van Loan, C. F. (2013).
Matrix computations, volume 3.
JHU press.

Gönen, M., Johnson, W. O., Lu, Y., and Westfall, P. H. (2005).
The Bayesian two-sample t test.
The American Statistician, 59(3):252–257.

Green, P. J. (1995).
Reversible jump markov chain monte carlo computation and Bayesian model determination.
Biometrika, 82(4):711–732.

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2020).
Scalable approximate mcmc algorithms for the horseshoe prior.
Journal of Machine Learning Research, 21(73):1–61.

References II

Kang, J., Reich, B. J., and Staicu, A.-M. (2018).
Scalar-on-image regression via the soft-thresholded Gaussian process.
Biometrika, 105(1):165–184.
code available at `https://www4.stat.ncsu.edu/~bjreich/software`.

Kim, Y. and Gao, C. (2020).
Bayesian model selection with graph structured sparsity.
Journal of Machine Learning Research, 21(109):1–61.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010).
Penalized regression, standard errors, and Bayesian lassos.
Bayesian Analysis, 5(2):369–411.

Liu, F., Chakraborty, S., Li, F., Liu, Y., and Lozano, A. C. (2014).
Bayesian regularization via graph laplacian.
Bayesian Analysis, 9(2):449–474.

Luo, Z. T., Sang, H., and Mallick, B. (2021).
A Bayesian contiguous partitioning method for learning clustered latent variables.
Journal of Machine Learning Research, 22(37):1–52.

Song, Q. and Cheng, G. (2020).
Bayesian fusion estimation via t shrinkage.
Sankhya A, 82(2):353–385.

References III

Teixeira, L. V., Assunção, R. M., and Loschi, R. H. (2019).
Bayesian space-time partitioning by sampling and pruning spanning trees.
Journal of Machine Learning Research, 20(85):1–35.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005).
Sparsity and smoothness via the fused lasso.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108.

Tibshirani, R. J., Taylor, J., et al. (2011).
The solution path of the generalized lasso.
The Annals of Statistics, 39(3):1335–1371.

Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. (2016).
Trend filtering on graphs.
Journal of Machine Learning Research, 17(105):1–41.

Yang, S., Yuan, L., Lai, Y.-C., Shen, X., Wonka, P., and Ye, J. (2012).
Feature grouping and selection over an undirected graph.
In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 922–930.