New Directions in Bayesian Shrinkage for Sparse and Structured Data

Jyotishka Datta

September 14, 2021

Virginia Tech

New Directions in Bayesian Shrinkage for Sparse and Structured Data

Part I: Global-Local Shrinkage: Overview

1. Sparse signal recovery
2. Horseshoe prior
3. Optimality properties
4. Global-local family

Part II: New Directions

1. Grouped sparsity/shrinkage
2. Precision matrix estimation
3. Future directions

Global-Local Shrinkage: A Brief Overview

Common Theme: High-dimensional Data

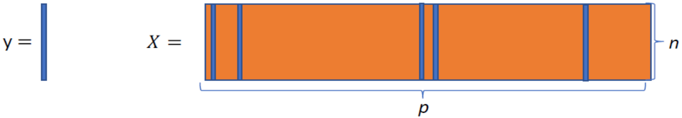
Sparsity: Needles in haystack !

High-dimensional Inference

Normal Means: $(Y_i \mid \theta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \sigma^2), i = 1, \ldots, n,$

Regression: $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},\ p \gg n,\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}).$

Sparsity: $\theta \in \ell_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}, p_n/n \to 0$



Theoretical Model

$$y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon$$

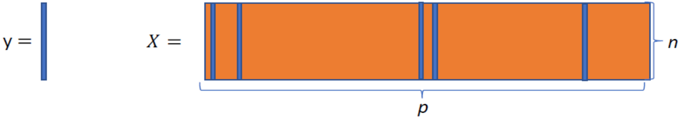
Fitted Model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots \hat{\beta}_p X_p$$

Normal Means: $(Y_i \mid \theta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \sigma^2), i = 1, \ldots, n,$

Regression: $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},\ p \gg n,\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$

Sparsity: $\theta \in \ell_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}, p_n / n \to 0$



Theoretical Model

$$y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon$$

Fitted Model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots \hat{\beta}_p X_p$$

Grouped covariates: $\mathbf{y} \sim \mathcal{N}(\boldsymbol{C\alpha} + \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g, \sigma^2 \boldsymbol{I}_n)$ where $g = 1, \ldots, G$ indexes the groups.

Precision matrix: $\mathbf{X}^{(n)} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, Estimate $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

Goals:

1. Recovery: provide estimator $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\Omega}}$.
2. Multiple Testing: Test whether each θ_i (or ω_{ij}) is zero or non-zero.
3. Variable selection.
4. Prediction.

The Two-groups Model i

- Natural hierarchical Bayesian solution : two-groups model.
  1. Assume each θ_i is non-zero with a prior probability π, and the non-zero θ_i's come from a common density $f_A(\cdot)$.
  2. Use Bayes' rule to calculate posterior probabilities that each $\theta_i \sim f_A(\cdot)$.
- Automatically adjusts for multiplicity and sparsity without any regularization.
- Carry out tests using the posterior inclusion probabilities (PIP).

$$\text{Posterior Inclusion Probability} = \omega_i = P(\theta_i \neq 0 \mid y_i)$$

- Induce sparsity through a 'spike and slab' prior.

The Two-groups Model ii

- Spike & Slab

$$Y_i \sim \mathcal{N}(\theta_i, \sigma^2),\ i = 1, \ldots, n$$

$$\theta_i \sim (1-p)\underbrace{\delta_{\{0\}}}_{\textit{Spike}} + p\overbrace{\mathcal{N}(0, \psi^2)}^{\textit{Slab}}$$

Multiple testing:

$$H_{0i} : \theta_i = 0 \text{ vs. } H_{Ai} : \theta_i \neq 0,\ i = 1, \ldots, n.$$

- Need (latent) indicators for MCMC:

$$\gamma_i = \begin{cases} 0 & \text{if } \theta_i = 0 \\ 1 & \text{if } \theta_i \neq 0 \end{cases}$$

- γ indexes $2^{\text{model dimension}}$ possible models: exploring the full posterior is computationally expensive.

Towards the one-group model i

- The two-groups model leads to a shrinkage rule linear in y_i.
- If $\theta_i \sim (1-p)\delta_{\{0\}} + p\mathcal{N}(0, \psi^2)$, the posterior mean is:

$$\mathbb{E}(\theta_i \mid y_i) = \omega_i \frac{\psi^2}{1+\psi^2} y_i = \omega_i^* y_i \tag{1}$$

  where ω_i is the posterior inclusion probability $P(\theta_i \neq 0 \mid y_i)$.
- If $\psi^2 \to \infty$ as the number of tests $n \to \infty$:

$$\boxed{E(\theta_i \mid y_i) \approx \omega_i y_i} \text{ (linear in } y_i\text{)}$$

- The one-group model takes a different route :
- *Directly models the posterior inclusion probability* ω_i

The One-group model

Global-local shrinkage priors: Horseshoe [Carvalho et al., 2010]

$$Y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2); \quad \theta_i \mid \lambda_i \sim \mathcal{N}\left(0, \lambda_i^2 \tau^2\right);$$

$$\underbrace{\lambda_i}_{\text{local}} \overset{\text{ind}}{\sim} \mathcal{C}^+(0,1), \underbrace{\tau}_{\text{global}} \sim \mathcal{C}^+(0,\sigma) \ \text{(Heavy-tailed prior)}$$

Posterior mean:
$\mathbb{E}(\theta_i \mid y_i) = \{1 - \mathbb{E}(1/1+\lambda_i^2\tau^2 \mid y_i)\} y_i \doteq (1 - \mathbb{E}(\kappa_i \mid y_i)) y_i$.

The One-group model

Global-local shrinkage priors: Horseshoe [Carvalho et al., 2010]

$$Y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2); \quad \theta_i \mid \lambda_i \sim \mathcal{N}\left(0, \lambda_i^2 \tau^2\right);$$

$$\underbrace{\lambda_i}_{\text{local}} \overset{\text{ind}}{\sim} \mathcal{C}^+(0,1), \underbrace{\tau}_{\text{global}} \sim \mathcal{C}^+(0,\sigma) \ \text{(Heavy-tailed prior)}$$

Posterior mean:

$\mathbb{E}(\theta_i \mid y_i) = \{1 - \mathbb{E}(1/1+\lambda_i^2\tau^2 \mid y_i)\}y_i \doteq (1 - \mathbb{E}(\kappa_i \mid y_i))y_i$.

Two-groups Model	One-group Model
$\mathbb{E}(\theta_i \mid y_i) \approx \omega_i y_i$, $\omega_i = \text{PIP}$	$\mathbb{E}(\theta_i \mid Y_i) = \{1 - \mathbb{E}(\kappa_i \mid y_i)\}y_i$

$1 - \mathbb{E}(\kappa_i \mid y_i)$ mimics the posterior inclusion probability ω_i.

$\mathbb{E}(\kappa_i \mid y_i) \approx 0$ for large y_i (signal), $\mathbb{E}(\kappa_i \mid y_i) \approx 1$ for small y_i (noise).

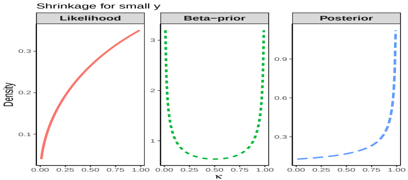
Why not use the two-groups model directly?

How to Build a Sparsity Prior

- $\mathbb{E}(\kappa_i \mid y_i) \approx 0$ for large y_i, $\mathbb{E}(\kappa_i \mid y_i) \approx 1$ for small y_i.

$$\kappa\text{-scale: } \underbrace{p(\kappa_i \mid y_i)}_{\text{posterior}} \propto \underbrace{p(y_i \mid \kappa_i)}_{\text{likelihood}} \underbrace{p(\kappa_i)}_{\text{prior}} \propto \boxed{\kappa_i^{\frac{1}{2}} \exp\left\{-\kappa_i \frac{y_i^2}{2}\right\}} p(\kappa_i)$$

- Likelihood doesn't concentrate near 1 for $y_i \approx 0$.
- Horseshoe: Push density towards 1 $\rightarrow$ replace $\kappa_i^{\frac{1}{2}}$ with $(1-\kappa_i)^{-\frac{1}{2}}$.
- Achieved by 'horseshoe': $p(\kappa_i) \propto 1/\sqrt{\kappa_i(1-\kappa_i)}$.



$$\lambda_i^2 \sim C^+(0,1) \equiv \kappa_i \sim \text{Be}(\tfrac{1}{2}, \tfrac{1}{2}) \Rightarrow \text{“Horseshoe”}.$$

Global-Local priors

Global-local scale mixtures[Polson and Scott, 2010b]:

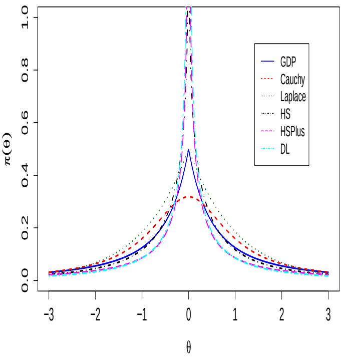
$$(\mathbf{y} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}); \; \theta_i \sim \mathcal{N}(0, \lambda_i^2 \tau^2)$$

$$\lambda_i^2 \sim \pi(\lambda_i^2); \; (\tau^2) \sim \pi(\tau^2), i = 1, \ldots, n.$$

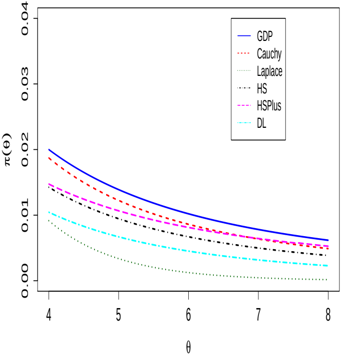
λ_i: local shrinkage - tags signal, τ: global shrinkage - adjusts to sparsity.

Global-local shrinkage priors	Authors
Normal Exponential Gamma	Griffin and Brown [2010]
Horseshoe	Carvalho et al. [2010, 2009]
Hypergeometric Inverted Beta	Polson and Scott [2010a]
Generalized Double Pareto	Armagan et al. [2011]
Generalized Beta	Armagan et al. [2013]
Dirichlet–Laplace	Bhattacharya et al. [2015]
Horseshoe+	Bhadra et al. [2017b]
Horseshoe-like	Bhadra et al. [2017a]
Spike-and-Slab Lasso	Ročková and George [2016]
R2-D2	Zhang et al. [2016]
Inverse-Gamma-Gamma	Bai and Ghosh [2017]
Heavy-tailed Horseshoe	Womack and Yang [2019]
Log-adjusted prior	Hamura et al. [2020]
Gauss–Hypergeometric	Datta and Dunson [2016]
Extremely heavy-tailed (EH) prior	Hamura et al. [2019]

Shape of G-L priors

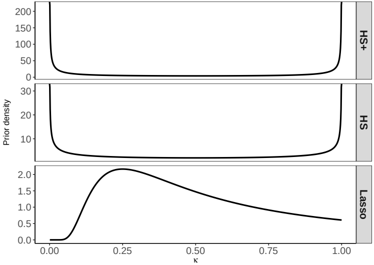


(a) Prior densities near origin

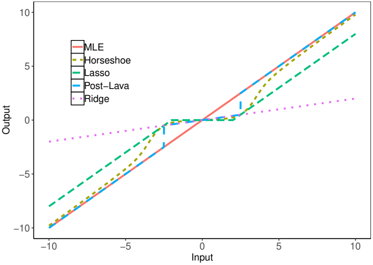


(b) Tails of prior densities

Need: Spike at zero and Heavy-tails



(a) Shrinkage profile for Horseshoe, Horseshoe+, and Laplace prior.



(b) Shrinkage Profiles

Lasso overshrinks, Horseshoe doesn't

Castillo et al. [2015]: the full Lasso posterior distribution does not contract **at the same speed as the posterior mode** $\Rightarrow$ Poor uncertainty quantification.

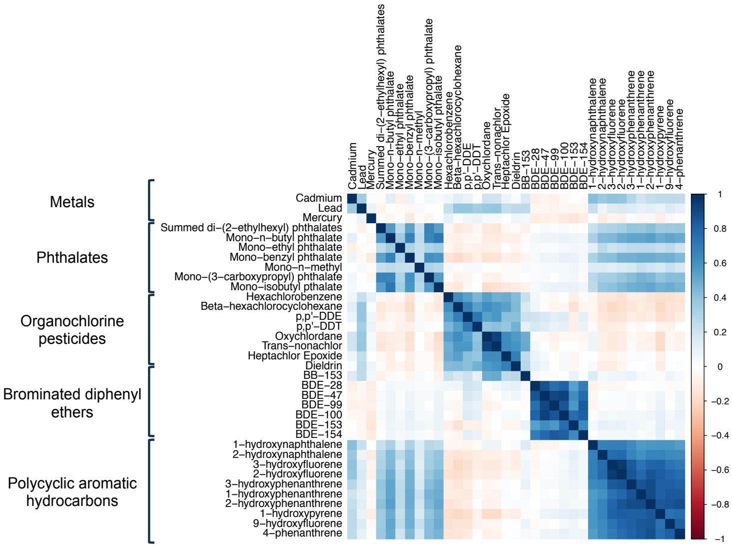
Theory for general G-L prior

$$\theta_i \sim \mathcal{N}(0, \lambda_i^2 \tau^2),\ \lambda_i^2 \overset{\text{ind}}{\sim} \pi_1(\lambda_i^2);\ (\tau^2) \sim \pi_2(\tau^2), i = 1, \ldots, n.$$

- Ghosh et al. [2016]: Bayes oracle for G-L priors.
- Ghosh and Chakrabarti [2017]: Asymptotic Minimaxity for G-L priors.
- Key idea: local shrinkage priors should have regularly varying tails.
- Up to $O(1)$ can be relaxed: G-L priors can be exactly minimax and ABOS [Ghosh et al., 2016, Bai and Ghosh, 2017].

Grouped shrinkage

Exposure Correlation Structure (NHANES 2003-2004)



Source: National Health and Nutrition Examination Survey (NHANES).

Simple multipollutant model

- Consider a Bayesian sparse linear regression model

$$[\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2] \sim N\left(\boldsymbol{C}\boldsymbol{\alpha} + \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g, \sigma^2 \boldsymbol{I}_n\right), \ \ \pi(\boldsymbol{\alpha}) \propto 1, \ \ \boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}), \tag{2}$$

  where $g = 1, \ldots, G$ indexes the groups, $\boldsymbol{y}$ is an $n \times 1$ vector of centered continuous responses, $\boldsymbol{C}$ is a matrix of adjustment covariates,
- and ... $\boldsymbol{X}_g$ is an $n \times p_g$ matrix of standardized covariates in the g-th group, $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gp_g})^\top$ is a $p_g \times 1$ vector of regression coefficients corresponding to the g-th group,
- and ... $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top$ is a $p \times 1$ vector of regression coefficients to employ shrinkage on.

Global-Group-Local Shrinkage Priors [Xu et al., 2016]

$$[y_{gj}|\beta_{gj}, \sigma^2] \sim N(\beta_{gj}, \sigma^2), \;\; [\beta_{gj}|\tau^2, \gamma_g^2, \lambda_{gj}^2] \sim N(0, \tau^2\gamma_g^2\lambda_{gj}^2),$$

where gj indexes the j-th mean in the g-th group, $\boldsymbol{\lambda}_g^2 = (\lambda_{g1}^2, ..., \lambda_{gp_g}^2)$, and p_g denotes the number of observations in the g-th group.

Key Idea: Need $\pi(\gamma_g^2, \boldsymbol{\lambda}_g^2)$ such that

$$\gamma_g^2\lambda_{gj}^2 \sim \beta'(a_g, b_g), \;\; \forall j \in \{1, ..., p_g\}.$$

Global-Group-Local Shrinkage Priors [Xu et al., 2016]

$$[y_{gj}|\beta_{gj}, \sigma^2] \sim N(\beta_{gj}, \sigma^2), \;\; [\beta_{gj}|\tau^2, \gamma_g^2, \lambda_{gj}^2] \sim N(0, \tau^2\gamma_g^2\lambda_{gj}^2),$$

where gj indexes the j-th mean in the g-th group, $\boldsymbol{\lambda}_g^2 = (\lambda_{g1}^2, ..., \lambda_{gp_g}^2)$, and p_g denotes the number of observations in the g-th group.

Key Idea: Need $\pi(\gamma_g^2, \boldsymbol{\lambda}_g^2)$ such that

$$\gamma_g^2\lambda_{gj}^2 \sim \beta'(a_g, b_g), \;\; \forall j \in \{1, ..., p_g\}.$$

Proposition: If $U \sim G(a, \eta)$ and $V \sim IG(b, \eta)$ are independent, then

$$UV \sim \beta'(a, b).$$

Group Inverse-Gamma Gamma (GIGG) Prior [Boss, Datta, Wang, Park, Kang, and Mukherjee, 2021]

Formulation

$$[\beta_{gj}|\tau^2, \gamma_g^2, \lambda_{gj}^2] \sim N(0, \tau^2\gamma_g^2\lambda_{gj}^2), \quad \gamma_g^2 \sim G(a_g, 1), \quad \lambda_{gj}^2 \sim IG(b_g, 1)$$

Here, the index gj refers to the j-th mean in the g-th group.

Posterior Distribution of Shrinkage Factors

$$\pi\left(\kappa_{g1}, ..., \kappa_{gp_g} | y_{g1}, ..., y_{gp_g}, \tau^2, \sigma^2, a_g, b_g\right) \propto$$

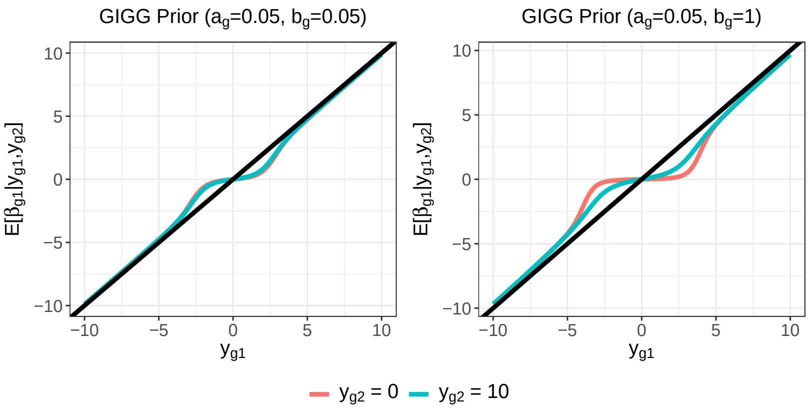
$$\left(1 + \frac{\tau^2}{\sigma^2} \sum_{j=1}^{p_g} \frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g + p_g b_g)} \left(\prod_{j=1}^{p_g} \kappa_{gj}^{b_g - 1/2} (1-\kappa_{gj})^{-(b_g+1)} e^{-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}}\right),$$

where $0 < \kappa_{gj} < 1$ for all $1 \le j \le p_g$ (p_g is the size of the g-th group).

Reduces to usual horseshoe prior for $p_g = 1$ (groups of size 1).

Posterior Mean (GIGG Prior)

Illustrative Model: $[y_{g1}|\beta_{g1}] \sim N(\beta_{g1}, 1), \;\; [y_{g2}|\beta_{g2}] \sim N(\beta_{g2}, 1)$



Here a_g effectively controls the overall strength of the shrinkage, whereas b_g generally controls the dependence of the within-group shrinkage.

Posterior Concentration (Sparse Normal Means)

- $|y_{gj}| \to \infty \implies$ posterior distribution of κ_{gj} concentrates near 0.
- $\tau \to 0 \implies$ posterior distribution of κ_{gj} concentrates near 1.

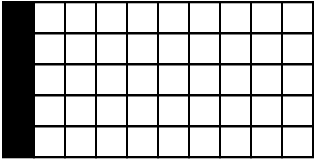
Posterior Concentration (Linear Regression with $p < n$)

- $\tau \to 0 \implies$ posterior distribution of $\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \cdot]\right\|_2$ concentrates near $\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2$ ($E[\boldsymbol{\beta} \mid \cdot]$ is the full conditional mean).
- For block diagonal correlation structure, $b_g \to \infty$ and τ^2/σ^2 small $\implies$ shrinkage of g-th group close to zero.

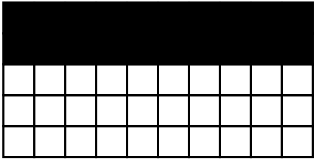
Posterior Consistency (Linear Regression)

- Assumes that $p = o(n)$ and fixed values of a_g and b_g.

Simulation Settings



(a) Concentrated Signal



(b) Distributed Signal

In the diagram, the gj-th box is the j-th exposure in the g-th group. The boxes corresponding to non-null regression coefficients are filled in.

Exposure Correlation Structure

- Correlations within exposure class are $\rho = 0.8$.
- Correlations between exposure classes are 0.2.

Mean-Squared Error

$\rho = 0.8$	Concentrated		Distributed	
Method	Null	Non-Null	Null	Non-Null
Ordinary Least Squares	3.74	0.41	8.09	2.03
Horseshoe	0.51	0.41	0.85	2.14
GIGG ($a_g = 1/n$, $b_g = 1/n$)	**0.11**	**0.30**	0.03	3.59
GIGG ($a_g = 1/2$, $b_g = 1/n$)	**0.11**	**0.30**	0.04	3.56
GIGG ($a_g = 1/n$, $b_g = 1/2$)	0.29	0.39	**0.03**	**1.57**
*GIGG ($a_g = 1/2$, $b_g = 1/2$)	0.33	0.40	0.24	1.70
GIGG ($a_g = 1/n$, $b_g = 1$)	0.53	0.49	**0.03**	**1.43**
GIGG ($a_g = 1/2$, $b_g = 1$)	0.58	0.49	0.26	1.43
GIGG (MMLE)	**0.20**	**0.34**	**0.04**	**1.42**
Group Half Cauchy	0.30	0.39	0.08	1.64
Spike-and-Slab Lasso	**0.15**	**0.33**	0.21	4.27
BGL-SS	2.01	0.80	**0.04**	**1.31**
BSGS-SS	0.23	0.42	0.04	1.84

*GIGG ($a_g = 1/2$, $b_g = 1/2$) is equivalent to group horseshoe.

**Bolded entries indicate the top four performers.

Illustrative Example from NHANES 2003-2004

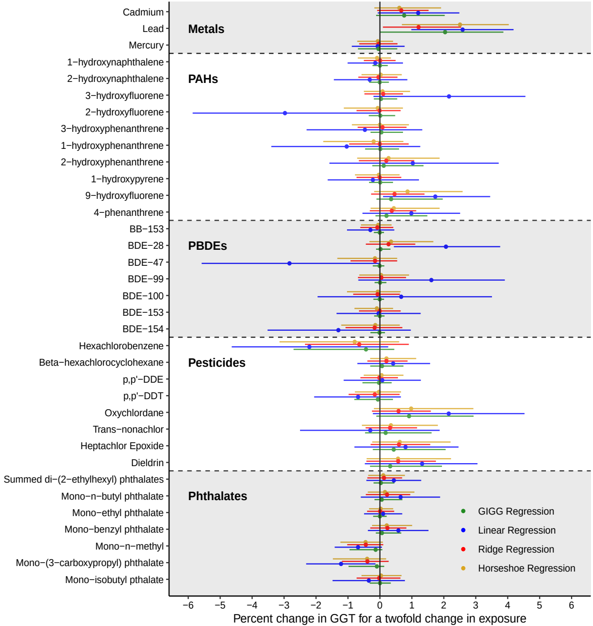
Study Details

- 990 adults with 35 measured environmental contaminants.
- Outcome of interest is Gamma-Glutamyl Transferase (GGT).

Exposure Classes

- 3 Metals (cadmium, lead, and mercury)
- 7 Phthalates
- 8 Organochlorine Pesticides
- 7 Polybrominated Diphenyl Ethers (PBDEs)
- 10 Polycyclic Aromatic Hydrocarbons (PAHs)

Illustrative Example from NHANES 2003-2004



Precision matrix estimation

- Gaussian graphical model (GGM) remains popular as a fundamental building block for network estimation because of the ease of interpretation of the resulting precision matrix estimate:
- An inferred off-diagonal zero corresponds to conditional independence of the two corresponding nodes given the rest [see, e.g., Lauritzen, 1996].
- There are both Bayesian and frequentist approaches to this, it is difficult to obtain good Bayesian and frequentist properties under the same prior–penalty dual, complicating justification.
- Our contribution is a **novel prior–penalty dual** that closely approximates the popular graphical horseshoe prior and penalty, and performs well in both Bayesian and frequentist senses.

- $\mathbf{X}^{(n)} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$.
- The corresponding precision matrix: $\boldsymbol{\Omega} = ((\omega_{ij}))$ is defined as $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.
- Assume that $\boldsymbol{\Omega}$ is sparse, in the sense that the number of non-zero off-diagonal elements is small.
- Goal: fully Bayesian inference on $\boldsymbol{\Omega}$, we need a suitable sparsity-favoring prior that also results in a penalty function with good frequentist properties.

Horseshoe Regularization [1]

- Horseshoe prior: $p(\omega)$ not analytically tractable !

$$\frac{1}{\tau(2\pi)^{3/2}} \log\left(1 + \frac{4\tau^2}{\omega^2}\right) < p_{HS}(\omega \mid \tau) < \frac{2}{\tau(2\pi)^{3/2}} \log\left(1 + \frac{2\tau^2}{\omega^2}\right),$$

- Hindrance in learning via EM-type algorithms.

[1] https://arxiv.org/abs/2104.10750

Horseshoe Regularization [1]

- Horseshoe prior: $p(\omega)$ not analytically tractable !

$$\frac{1}{\tau(2\pi)^{3/2}} \log\left(1 + \frac{4\tau^2}{\omega^2}\right) < p_{HS}(\omega \mid \tau) < \frac{2}{\tau(2\pi)^{3/2}} \log\left(1 + \frac{2\tau^2}{\omega^2}\right),$$

- Hindrance in learning via EM-type algorithms.
- Solution: normalize the tight bounds: 'horseshoe-like' [Bhadra et al., 2017a].

$$p_{\widetilde{HS}}(\omega \mid a) = \frac{1}{2\pi a^{1/2}} \log\left(1 + \frac{a}{\omega^2}\right).$$

- Extend this for precision matrix estimation (Sagar, Banerjee, D., & Bhadra, 2021).

[1] https://arxiv.org/abs/2104.10750

Normal Scale Mixture Representation!

- Frullani's identity [Jeffreys and Swirles, 1972, pp. 406–407]:

$$\int_0^\infty \frac{f(ax) - f(bx)}{x} dx = \{f(0) - f(\infty)\} \log(b/a),$$

- $f(x) = \exp(-x)$ yields a latent variable representation:

$$\frac{1}{2\pi a^{1/2}} \log\left(1 + \frac{a}{\omega^2}\right) = \int_0^\infty \exp\left(-\frac{u\omega^2}{a}\right) \frac{(1 - e^{-u})}{2\pi a^{1/2} u} du$$

- Normal scale mixture:

$$(\omega \mid u, a) \overset{ind}{\sim} \mathcal{N}\left(0, \frac{a}{2u}\right),\ p(u) = \frac{1 - e^{-u}}{2\pi^{1/2} u^{3/2}}$$

- Reparametrize ($t^2 = 2u$ and $\tau^2 = a$):

$$(\omega \mid i, \tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{t^2}\right), p(t) = \frac{(1 - e^{-\frac{1}{2}t^2})}{\sqrt{2\pi} t^2}$$

- This $p(t)$ is the standard Slash-Normal density that can be written as a Normal variance mixture with a $\mathrm{Pareto}(\frac{1}{2})$ mixing density.

Details

- For the fully Bayesian model, the element-wise prior specification induced by the horseshoe-like prior is,

$$\omega_{ij} \mid a \quad \sim \quad \pi(\omega_{ij} \mid a),\ 1 \leq i < j \leq p; \qquad \omega_{ii} \propto 1,\ 1 \leq i \leq p,$$

  where $\pi(\omega_{ij} \mid a)$ is the density of the horseshoe-like prior.
- The horseshoe-like prior above can be expressed as a Gaussian scale-mixture [Bhadra et al., 2017a], thus giving us a global-local shrinkage prior:

$$\omega_{ij} \mid \nu_{ij}, a \sim \mathcal{N}\left(0, \frac{a}{2\nu_{ij}}\right),\ \pi(\nu_{ij}) \sim \frac{1 - \exp(-\nu_{ij})}{2\pi^{1/2}\nu_{ij}^{3/2}}. \qquad (3)$$

- Only ν_{ij} is considered to be latent and the global scale parameter a is considered to be fixed.
- Can estimate a by the effective model size approach of Piironen et al. [2017] to avoid it collapsing to zero.

Details

- We restrict the prior on a subspace of symmetric positive definite matrices, $\mathcal{M}_p^+(L)$, where

$$\mathcal{M}_p^+(L) = \{\boldsymbol{\Omega} \in \mathcal{M}_p^+ : 0 < L^{-1} \leq \mathrm{eig}_1(\boldsymbol{\Omega}) \leq \cdots \leq \mathrm{eig}_p(\boldsymbol{\Omega}) \leq L < \infty\}. \tag{4}$$

- Only necessary for arriving at the theoretical results involving the posterior convergence rate of $\boldsymbol{\Omega}$. We assume that L is a fixed constant, which can be large.
- However, this condition does not affect the practical implementation of our proposed method, and is used purely as a technical requirement.
- Beyond this, no structural assumption (e.g., decomposability) is placed on either $\boldsymbol{\Omega}$ or $\boldsymbol{\Sigma}$.

Joint prior

- Combining the unrestricted prior as in (3) and (3), along with the prior space restriction as in (4), the joint prior distribution on $\mathbf{\Omega}$ is given by,

$$\pi(\mathbf{\Omega} \mid \nu, a)\pi(\nu) \propto \prod_{i,j:i<j} (1 - \exp(-\nu_{ij}))\, \nu_{ij}^{-1} \exp\left(\frac{-\nu_{ij}\omega_{ij}^2}{a}\right) \mathbb{1}_{\mathcal{M}_p^+(L)}(\mathbf{\Omega}). \tag{5}$$

- With the prior specification as in (5), the log-posterior $\mathcal{L}$ thus becomes,

$$\mathcal{L} \propto \frac{n}{2}\log|\mathbf{\Omega}| - \frac{n}{2}\mathrm{tr}(\mathbf{S}\mathbf{\Omega}) + \sum_{i,j:i<j}\left\{\log\left(1 - \exp(-\nu_{ij})\right) - \log\nu_{ij} - \frac{\nu_{ij}\omega_{ij}^2}{a}\right\} \tag{6}$$

- Utilize the Gaussian mixture representation to devise an Expectation Conditional Maximization (ECM) [Meng and Rubin, 1993] approach to MAP estimation.
- For updating the elements of the precision matrix, we use the coordinate descent technique proposed by Wang [2014].
- **E Step:** Following Bhadra et al. [2017a], we calculate the conditional expectation of the latent variable $\nu_{ij}, 1 \leq i < j \leq p$, at current iteration (t) as follows:

$$\mathbb{E}(\nu_{ij} \mid \omega_{ij}^{(t)}, a) = \left\{ \log(1 + \frac{a}{(\omega_{ij}^{(t)})^2}) \right\}^{-1} \frac{a^2}{((\omega_{ij}^{(t)})^2 + a)((\omega_{ij}^{(t)})^2)}. \tag{7}$$

- **CM Steps:** Having updated the latent parameters in the E-Step, the coordinate descent approach of Wang [2014] is used to update one column of the precision matrix at a time.

Posterior sampling

- Posterior sampling strategy combines ideas from [Bhadra et al., 2017a] and [Li et al., 2017].
- With substitutions $2\nu_{ij} \mapsto t_{ij}^2$ and $a \mapsto \tau^2$, the prior can be written as:

$$\omega_{ij} \mid \nu_{ij}, \tau \sim \mathcal{N}\left(0, \tau^2 / t_{ij}^2\right), \ \pi(t_{ij}) = \frac{1 - \exp\left(-t_{ij}^2/2\right)}{(2\pi)^{1/2} t_{ij}^2}, \ t_{ij} \in \mathbb{R}, \ \tau^2 > 0,$$

  where $\pi(t_{ij})$ above is known as the the slash normal density, expressed as $(\phi(0) - \phi(t_{ij}))/t_{ij}^2$ [Bhadra et al., 2017a].
- Introducing a further local latent variable r_{ij}, the density for t_{ij} can also be written as a normal scale mixture, where the scale follows a Pareto distribution, that is,

$$t_{ij} \mid r_{ij} \sim \mathcal{N}(0, r_{ij}), \ r_{ij} \sim \text{Pareto}\,(1/2)\,.$$

- Remaining steps are similar to the graphical horseshoe sampler of Li et al. [2017].

Posterior consistency

- Posterior contraction rate of the precision matrix $\mathbf{\Omega}$ around the true precision matrix $\mathbf{\Omega}_0$ with respect to the Frobenius norm under the graphical horseshoe-like prior.
- We make certain assumptions on the true precision matrix, the dimension and sparsity, and the prior space.
- **Assumptions:** True underlying graph is sparse, effective dimension of the parameter $\mathbf{\Omega}_0$, $p+s$ satisfies $(p+s)\log p/n = o(1)$, the prior space contains the true precision matrix, and the prior puts sufficient mass around the true zero elements in the precision matrix.

Theorem
The posterior distribution of $\mathbf{\Omega}$ satisfies

$$\mathbb{E}_0\left[P\{\|\mathbf{\Omega}-\mathbf{\Omega}_0\|_2 > M\epsilon_n \mid \mathbf{X}^{(n)}\}\right] \to 0,$$

for $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ and a sufficiently large constant $M > 0$.

MAP estimator

- We can prove that the extended real-valued penalty function $pen_a(x) = -\log\log(1 + a/x^2),\ a > 0$, is strongly concave, and hence strictly concave, for all $x \in dom(pen_a)$, separately for $x > 0$ and $x < 0$.
- Strict concavity of penalty function guarantees that the LLA algorithm will satisfy an ascent property, that is, $Q(\mathbf{\Omega}^{(t+1)}) > Q(\mathbf{\Omega}^{(t)})$.

Theorem

Under the conditions of Theorem 1, the MAP estimator of $\mathbf{\Omega}$, given by $\hat{\mathbf{\Omega}}^{\mathrm{MAP}}$ is consistent, in the sense that

$$\|\hat{\mathbf{\Omega}}^{\mathrm{MAP}} - \mathbf{\Omega}_0\|_2 = O_P(\epsilon_n),$$

where ϵ_n is the posterior convergence rate as defined in Theorem 1.

- Converges to the true precision matrix $\mathbf{\Omega}_0$ at the same rate as the posterior convergence rate in the Frobenius norm.

Hubs. The rows/columns are partitioned into K disjoint groups $G_1, \dots, G_K$. The off-diagonal entries ω_{ij}^0 are set to 0.25 if $i \neq j$ and $i, j \in G_k$ for $k = 1, \dots, K$. In our simulations we consider $p/10$ groups with equal number of elements in each group.

Table 1: 50 data sets generated with precision matrix $\mathbf{\Omega}_0$, where $n = 120$ and $p = 100$. Candidates: frequentist graphical lasso with penalized diagonal elements (GL1) and with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), the graphical horseshoe (GHS), graphical horseshoe-like ECM (ECM) and graphical horseshoe-like MCMC (MCMC).

	Hubs 90 nonzero pairs out of 4950 nonzero elements = 0.25						
	GL1	GL2	GSCAD	BGL	GHS	ECM	MCMC
Stein's loss	5.255	6.328	5.213	43.042	5.101	**4.22**	5.310
	(0.263)	(0.414)	(0.261)	(0.802)	(0.455)	(0.369)	(0.485)
F norm	3.018	3.432	3.003	4.295	2.544	**2.415**	2.687
	(0.091)	(0.112)	(0.093)	(0.156)	(0.126)	(0.103)	(0.141)
TPR	.995	.986	**.998**	.995	.872	0.985	0.754
	(.007)	(.017)	(.002)	(.008)	(.04)	(.014)	(0.004)
FPR	.101	.045	.983	.186	**.003**	.062	**0.003**
	(.016)	(.008)	(.012)	(.007)	(.001)	(0.005)	(0.001)
MCC	0.373	0.523	0.016	0.27	**0.85**	0.458	0.775
	(.027)	(.039)	(.006)	(.006)	(.027)	(.015)	(.033)
Avg CPU time	1.739	1.76	48.54	549.196	252.94	5.811	537.604

Summary and Scopes (Part I)

- Global-local priors: state-of-the-art Bayesian tool for sparse signal recovery.
- Can be extended to sparse + structured covariates: GIGG and graphical-horseshoe.
- Scale mixture: allows for MCMC + EM and LLA algorithms.
- Can be interpreted as non-convex penalty (horseshoe-like)
- Scopes:
  1. Selection for bi-level sparsity (still Oracle?)
  2. Multiple graphical models.
  3. Extend beyond Gaussian set-up (e.g. [Datta and Dunson, 2016]).
  4. An appealing new direction is Bayesian neural net, e.g. [Ghosh and Doshi-Velez, 2017] [‘Model selection in Bayesian neural networks via horseshoe priors’]

References

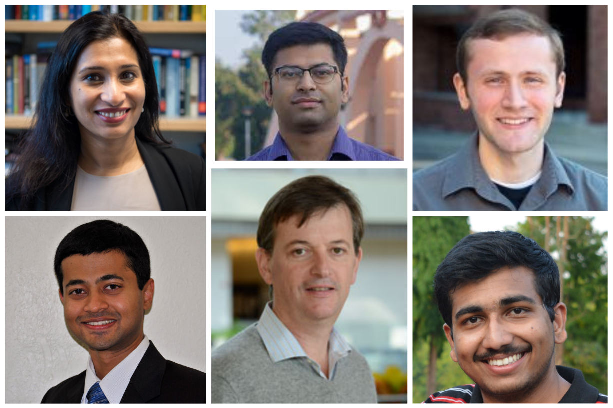
References (for this talk)

- **Graphical horseshoe-like prior:** Sagar, Ksheera, Banerjee, S., **Datta, J**., and Bhadra, A.. "Precision Matrix Estimation under the Horseshoe-like Prior-Penalty Dual." arXiv preprint arXiv:2104.10750 (2021).
- **GIGG shrinkage**: Boss, J., **Datta, J.**, Wang, X., Park, S. K., Kang, J., & Mukherjee, B. (2021). Group Inverse-Gamma Gamma Shrinkage for Sparse Regression with Block-Correlated Predictors. arXiv preprint arXiv:2102.10670.
- **Horseshoe-like prior**: Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. (2019). Horseshoe regularization for feature subset selection. *Sankhya B*. [preprint]
- **Graphical horseshoe:** Li, Y., Craig, B. A., & Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. Journal of Computational and Graphical Statistics, 28(3), 747-757.

References (General global-Local)

- Bhadra, A., **Datta, J.**, Li, Y., Polson, N. G., & Willard, B. (2019). Prediction risk for global-local shrinkage regression. **20 (78)**, 1-39, Journal of Machine Learning Research. arXiv:1605.04796.
- Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. T. (2019). Lasso Meets Horseshoe: A Survey. **34(3)**, 405-427. Statistical Science.
- Bhadra, **Datta**, Li and Polson (2019). "Horseshoe Regularization for Machine Learning in Complex and Deep Models". *Published*, *International Statistical Review. Discussed paper* [preprint].
- Bhadra, **Datta**, Polson, and Willard (2019), (*alphabetical), "Global-local mixtures - A Unifying Framework".*Accepted*, *Sankhya A*.
- Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis, 12(4), 1105-1131.
- **Datta, J.**, & Dunson, D. B. (2016). Bayesian inference on quasi-sparse count data. Biometrika, 103(4), 971-983.
- Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. Biometrika, 103(4), 955-969.
- **Datta, J.**, & Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. Bayesian Analysis, 8(1), 111-132.
- Li, **Datta**, Craig, and Bhadra, (2020+). "Joint Mean-Covariance Estimation via the Horseshoe with an Application in Genomic Data Analysis". *submitted*. [preprint].

Thank you!



Resources for Horseshoe Prior

Learning τ

1. Maximum marginal likelihood estimator (MMLE)
2. Full Bayes estimator: half-Cauchy prior truncated to the interval $[1/n, 1]$.
3. Cross-validation.
4. By studying the prior for $m_{\text{eff}} = \sum_{i=1}^n (1 - \kappa_i)$ [Piironen and Vehtari, 2016]

- MMLE beats simple thresholding:

$$\hat{\tau}_s(c_1, c_2) = \max \left\{ \frac{\sum_{i=1}^n \mathbf{1}\{|y_i| \geq \sqrt{c_1 \log(n)}\}}{c_2 n}, \frac{1}{n} \right\} .$$

- Empirical Bayes estimate of τ can replace a full Bayes estimate of τ.
- Caution to prevent the estimator from getting too close to zero.

Computation for Horseshoe

1. MCMC : block-updating $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ and τ using either a Gibbs or parameter expansion or slice sampling strategy.
2. Makalic and Schmidt [2016]: Inverse-gamma scale mixture for Gibbs sampling scheme for horseshoe and horseshoe+ prior for linear regression and logistic and negative binomial regression.
3. Hahn et al. [2016]: Elliptical slice sampler – wins over Gibbs strategies!
4. Bhattacharya et al. [2016]: Gaussian sampling alternative to the naïve Cholesky decomposition to reduce the computational burden from $O(p^3)$ to $O(n^2 p)$.

Implementation

Table 2: Implementations of Horseshoe and Other Shrinkage Priors

Implementation (Package/URL)	Authors
R package: `monomvn`	Gramacy and Pantaleo [2010]
R code in paper	Scott [2010]
R package: `horseshoe`	van der Pas et al. [2016]
R package: `fastHorseshoe`	Hahn et al. [2016]
MATLAB code	Bhattacharya et al. [2016]
GPU accelerated Gibbs sampling	Terenin et al. [2016]
`bayesreg` + MATLAB code in paper	Makalic and Schmidt [2016]
MATLAB code	Johndrow and Orenstein [2017]