Global-Local Shrinkage Priors: An Overview and New Directions

Jyotishka Datta

October 25, 2024

Virginia Tech

Outline of My Talk

Global-Local Shrinkage Priors: An Overview and New Directions

Part I: Grouped covariates

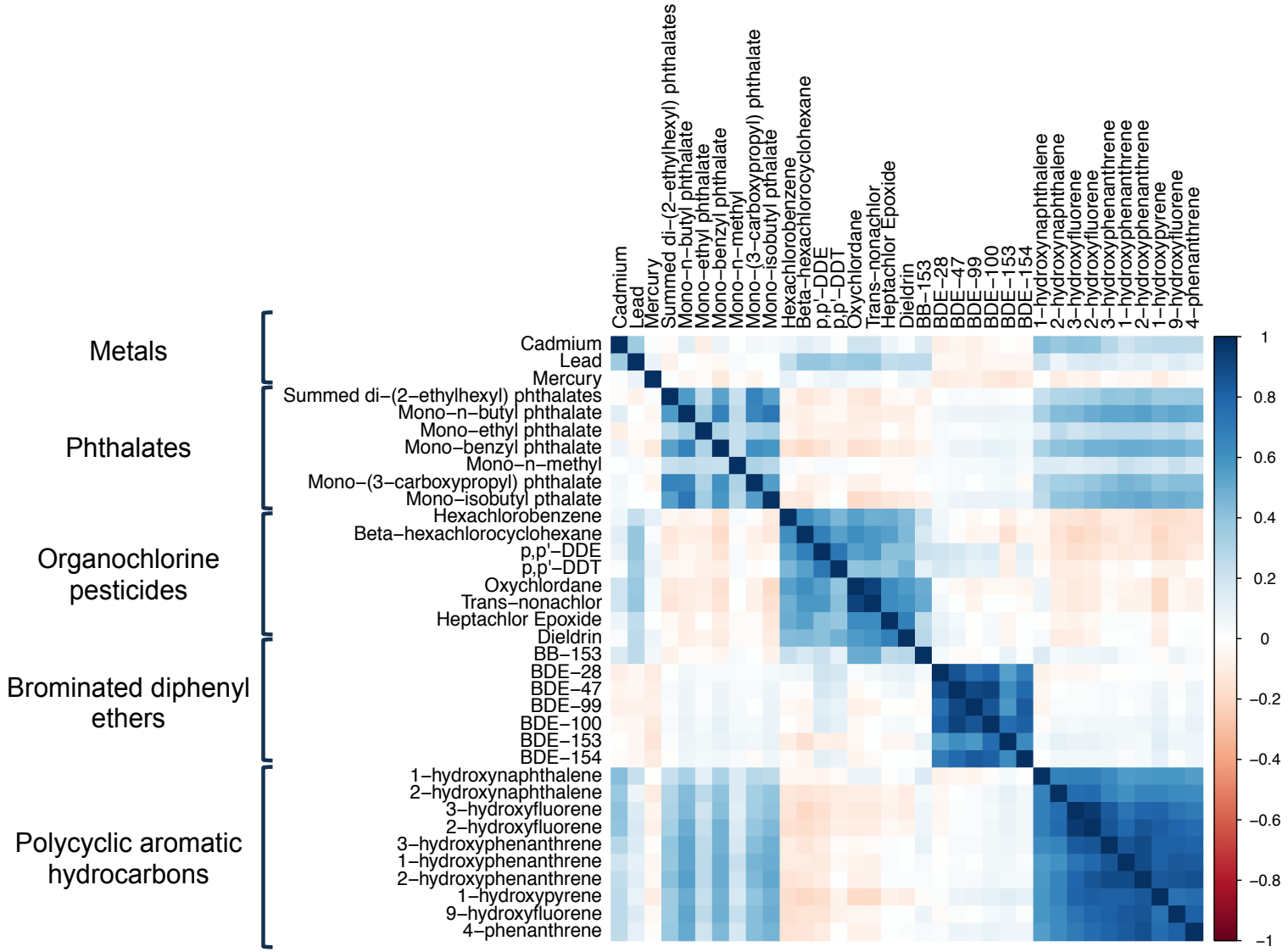
1. Global-Local Shrinkage: Overview
2. Global-local family
3. **Grouped covariates**
4. **GIGG prior**
5. Properties and application

Part II: Shrinkage on simplex

1. **Quasi-sparse count**
2. **Compositional data**
3. *Sparse generalized* Dirichlet prior
4. Future directions

Motivation

Example I: Exposure Correlation Structure (NHANES 2003-2004)



Example II: Rare Mutations

- Aim 1: Identify where rare mutations converge.
- Aim 2: Which ones are harmful?

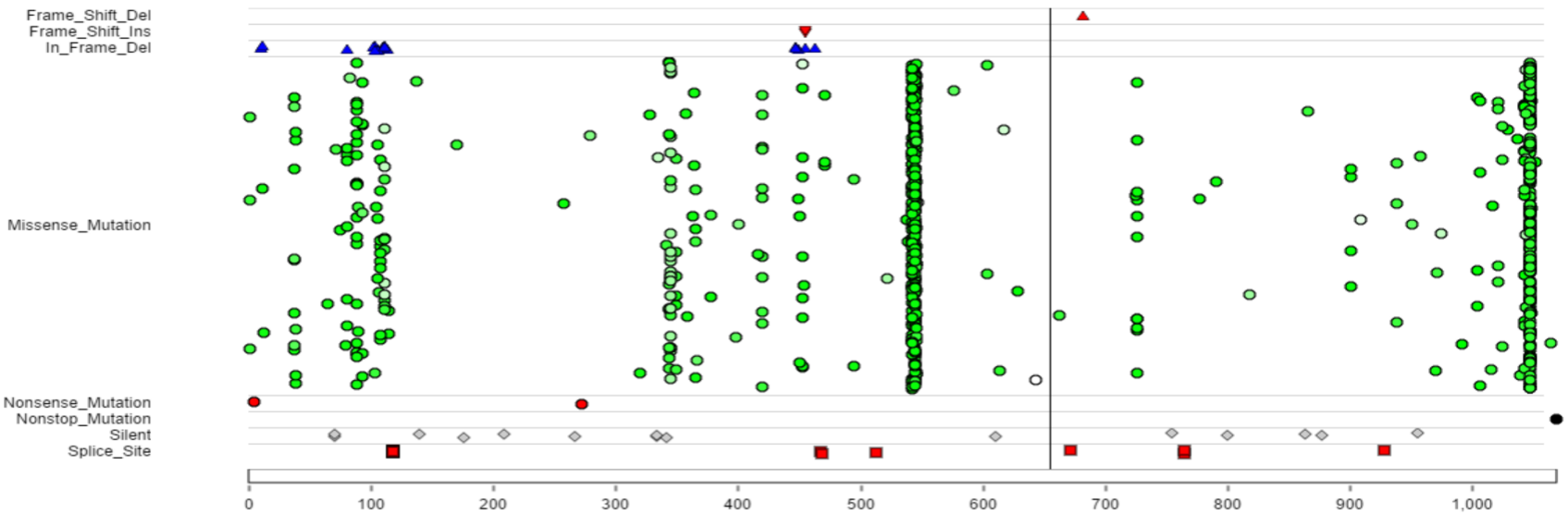
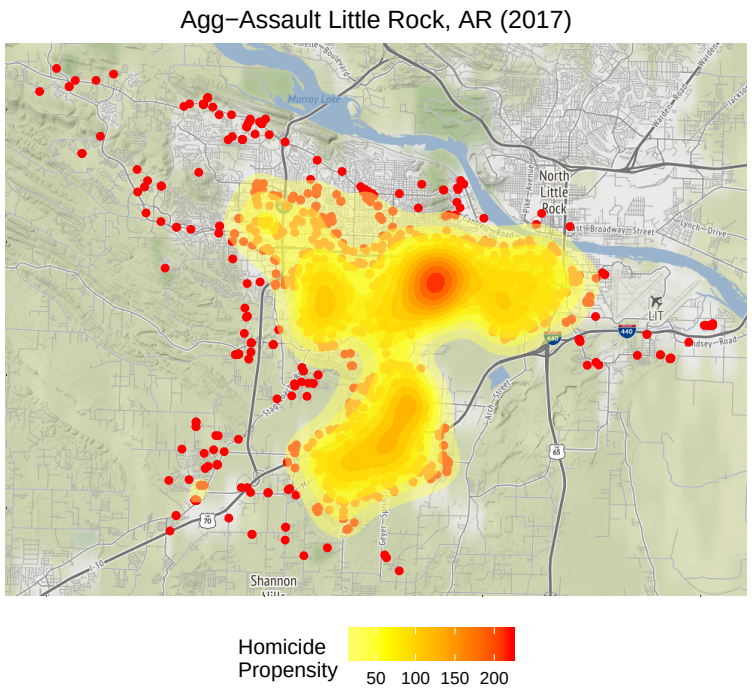
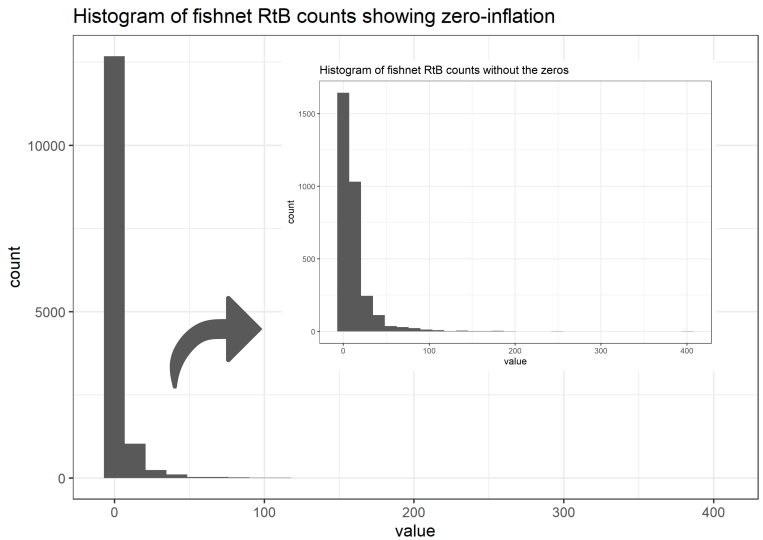


Figure 1: Rare Mutations on the oncogene 'PIK3CA'

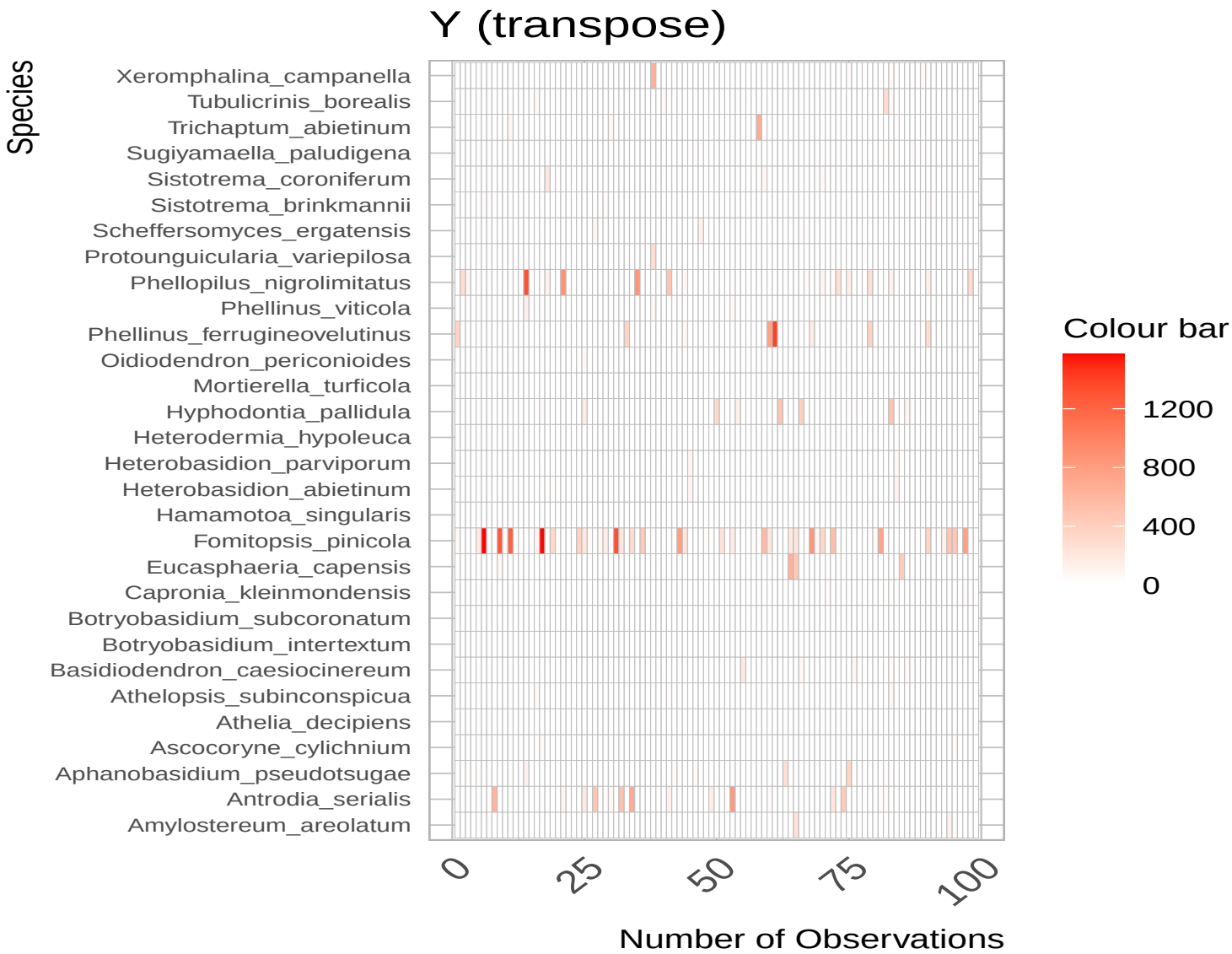
Example III: Crime forecasting

- Count data: over-abundance of zeros and near-zeros + sample size heterogeneity.
- Crime data: predict + effect of built environment





Example IV: Fungal biodiversity data



Source: [Ovaskainen et al., 2013] (species present in ≥ 10 logs).

Canonical Problems

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

Sparsity: $\theta \in \ell_0[s_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq s_n\}, s_n \ll n.$

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

Theoretical Model

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

Fitted Model

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

New Directions

Grouped covariates: $\mathbf{y} = \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}$ where $g = 1, \ldots, G$ indexes the groups. (Pollutants.)

Quasi-sparse counts: $(Y_i \mid \theta_i) \overset{\text{ind}}{\sim} \text{Poi}(\theta_i), i = 1, \ldots, n$. (Rare mutations, crime.)

Compositional: $\mathbf{y} \sim \text{Multinomial}(n; (\pi_1, \ldots, \pi_K))$, K categories. (Fungi, Gut microbiome.)

Goals:

Recovery, Model selection, Prediction, or Structure learning.

Global-Local Shrinkage: A Brief Overview

Continuous shrinkage priors

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

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

Continuous shrinkage priors

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

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
$\mathbb{E}(\theta_i \mid y_i) \approx \omega_i y_i$, $\omega_i =$ 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).

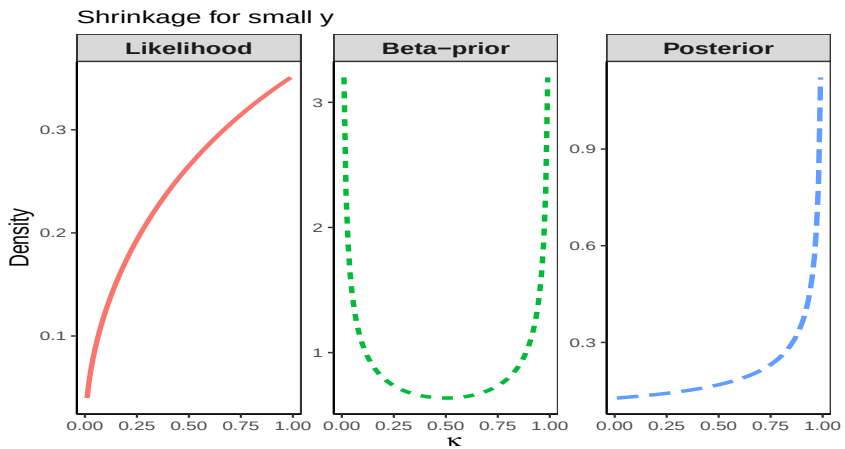
Why not use the two-groups model directly?

How to build a sparsity prior?

- Once again, consider the sparse normal means problem.
- $\mathbb{E}(\kappa_i \mid y_i) \approx 0$ for large y_i, $\mathbb{E}(\kappa_i \mid y_i) \approx 1$ for small y_i.

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

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

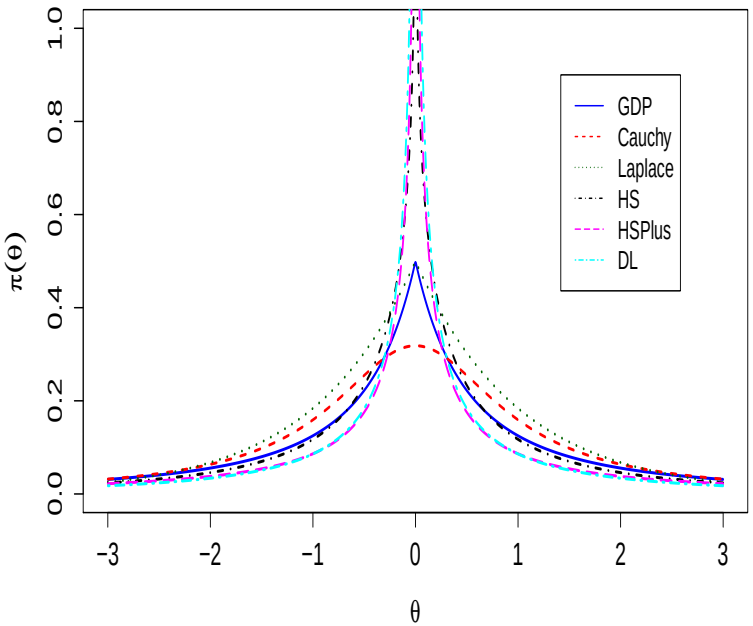
Global-local scale mixtures[Polson and Scott, 2010b]:

$$(\mathbf{y} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I});\ \theta_i \sim \mathcal{N}(0, \lambda_i^2\tau^2)$$
$$\lambda_i^2 \sim \pi(\lambda_i^2);\ (\tau^2) \sim \pi(\tau^2), i = 1, \ldots, n.$$

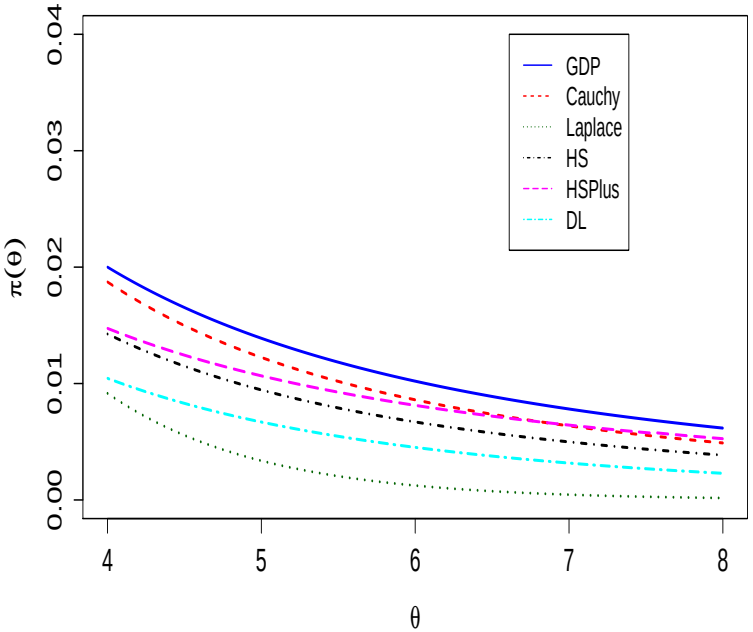
λ_i: local shrinkage - tags signal, τ: global shrinkage - adjusts to sparsity.

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]
Regularized Horseshoe	Piironen et al. [2017]
Gauss–Hypergeometric	Datta and Dunson [2016]
Extremely heavy-tailed (EH) prior	Hamura et al. [2021]

Shape of G-L priors



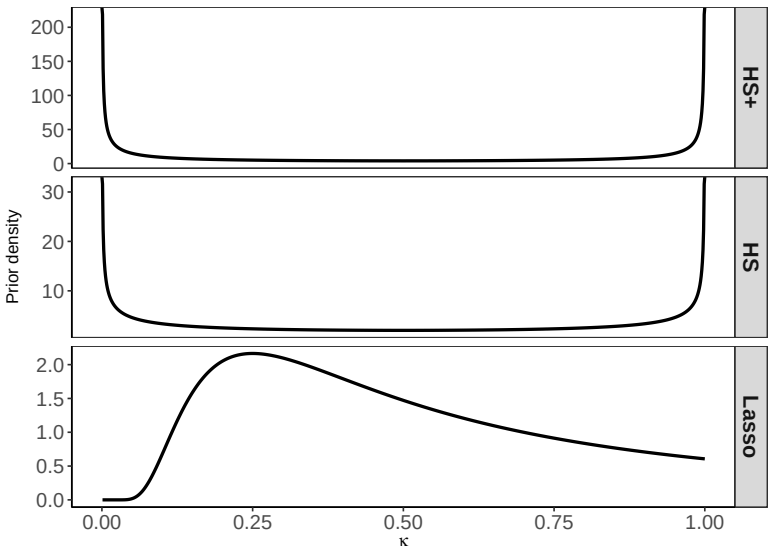
(a) Prior densities near origin



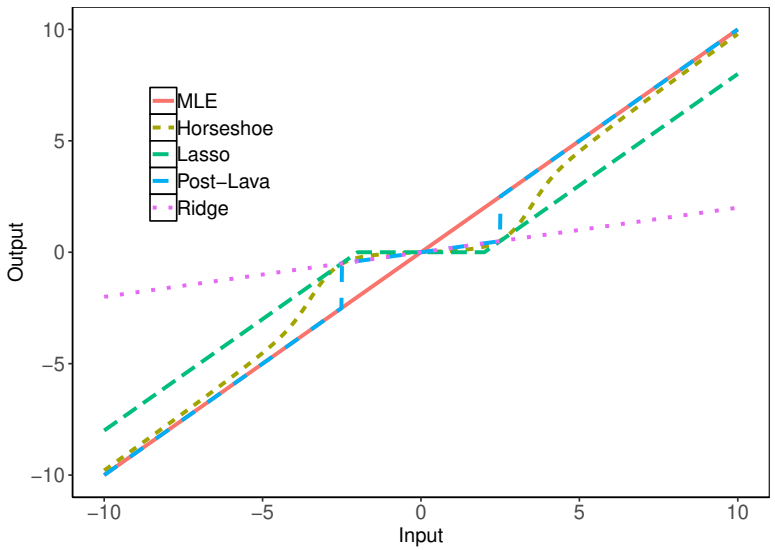
(b) Tails of prior densities

Need: Spike at zero and Heavy-tails

Horseshoe vs. (Bayesian) Lasso



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



(b) Shrinkage Profiles

Castillo et al. [2015]: the full Lasso posterior distribution does not contract **at the same speed as the posterior mode** $\Rightarrow$ Poor uncertainty quantification[1].

[1] See Castillo and van der Vaart [2012] for a similar phenomenon for the hard-spike-and-slab prior and ℓ_q loss ($0 < q < 1$).

Theory for G-L prior

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

- Variable selection: Asymptotic Bayes oracle [Datta and Ghosh, 2013, Ghosh et al., 2016].
- *Bayes oracle*: Risk for oracle = lower bound of $(1/n)$ times the risk for any multiple testing procedure under the two-groups model.
- Estimation: near-minimaxity [van der Pas et al., 2014, 2017].

$$\sup_{\boldsymbol{\theta} \in \ell_0[s_n]} \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}} \|\hat{\theta}_{HS}(\mathbf{y}) - \boldsymbol{\theta}\|^2 \asymp s_n \log(n/s_n), \tag{1}$$

- Better **finite sample** predictive risk using Stein's unbiased risk estimate (SURE framework): [Bhadra et al., 2016a].
- Key idea: local shrinkage priors should have **regularly varying** tails.

Challenges

Collapsing behaviour of τ

- HS posterior mean could collapse to the null $\hat{\boldsymbol{\beta}} \approx 0$ for **moderately sparse** regime $\Longrightarrow$ poor performance.
- Solutions: Bayes $\sqrt{\text{Lasso}}$ [Abba, 2018], GLT prior [Lee et al., 2020], Inverse-Gamma Gamma [Bai and Ghosh, 2019].
- Full Bayes estimator: half-Cauchy prior truncated to the interval $[1/n, 1]$.
- Effective model size approach: $m_{\text{eff}} = \sum_{i=1}^{n}(1 - \kappa_i)$ [Piironen et al., 2017].

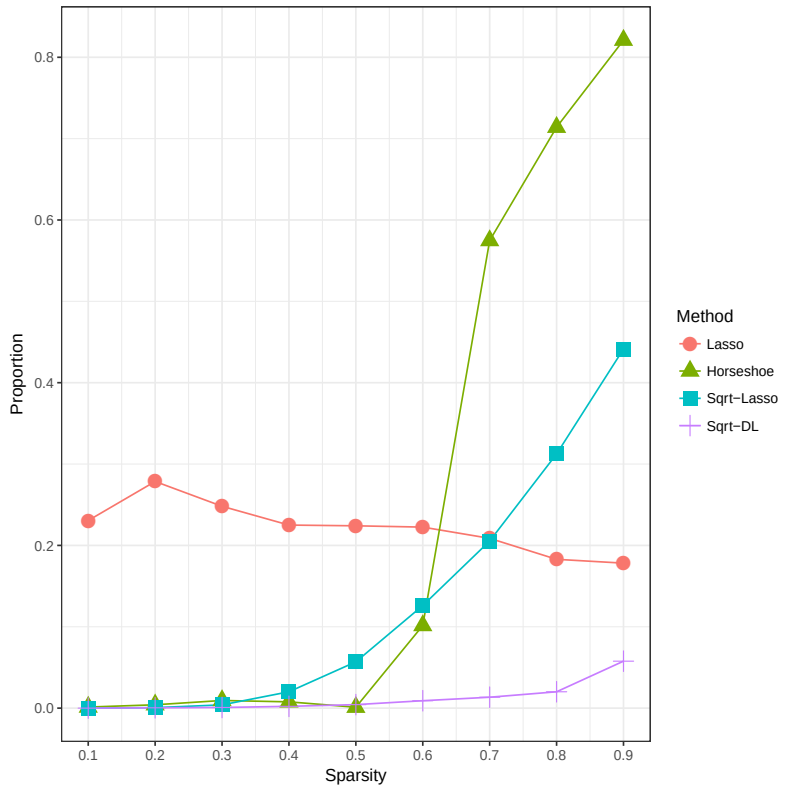


Figure 5: Misclassification probability for Horseshoe in sparse to dense regime

Computation for Horseshoe

1. Suggesting continuous shrinkage priors are computationally superior to spike-and-slab is somewhat misleading.
2. MCMC algorithms for continuous shrinkage priors can have poor mixing behavior.
3. Usual idea: block-updating $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ and τ using either a Gibbs or parameter expansion or slice sampling strategy.
4. Makalic and Schmidt [2016]: Inverse-gamma scale mixture for Gibbs sampling scheme for linear/logistic/negative binomial regression.
5. Hahn et al. [2016]: elliptical slice sampler – wins over Gibbs strategies!
6. Bhattacharya et al. [2016]: Gaussian sampling alternative to the naïve Cholesky decomposition: computational burden from $O(p^3)$ to $O(n^2 p)$.
7. Johndrow et al. [2020]: *Approximate* MCMC based on thresholding (GWAS scale, p $100k$).

UQ for Horseshoe

- The marginal credible intervals have asymptotically correct frequentist coverage [van der Pas et al., 2017] for parameters that are either very close to zero or above the detection threshold.
- Signals in a certain 'intermediate' range are shrunk too much toward zero for credible intervals to have correct coverage.
- Baraud [2002], Li [1989] and others: Impossibility theorem: honesty vs. adaptivity. Need extra conditions such as 'Excessive Bias Restriction.'

Recent developments i

- Horseshoe prior can be sharpened if local shrinkage priors have heavier tails. [Bhadra, Datta, Polson, Willard, et al., 2017b, Bayes Analysis].
- Horseshoe priors can be used for non-linear problems, e.g. $\psi(\boldsymbol{\theta}) = \sum_{i=1}^{n} \theta_i^2$ - resolves marginalization paradoxes [Bhadra et al., 2016b, Bka].
- Horseshoe has better **finite sample** predictive risk over alternative shrinkage methods (including Lasso, ridge, PCR and sparse PLS) using Stein's unbiased risk estimate (SURE framework) [Bhadra et al., 2016a].
- We can build shrinkage priors for quasi-sparse count data: Gauss-hypergeometric prior [Datta and Dunson, 2016, Bka].
- Horseshoe prior can be closely approximated by another scale mixture prior horseshoe-like with close analytic form, leading to non-convex penalty as well as full Bayes [Bhadra et al., 2017a].

Recent developments ii

- Gaussian graphical model: use shrinkage priors on the off-diagonal elements of the precision matrix Ω [Sagar Ksheera et al., 2021]
- Bi-level sparsity: a new prior, GIGG, is designed explicitly for predictors with bi-level grouping structure [Boss et al., 2021]

Grouped shrinkage

Simple multipollutant model

- Sparse Linear Models with Block-Correlated Regressors [2]

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

  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 ... $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top$ is a $p \times 1$ vector of regression coefficients to employ shrinkage on.
- Bi-level sparsity: not all groups are important, and within each group not all effects are important.

[2] Joint work with Jonathan Boss, Bhramar Mukherjee, Xin Wang, Sung Kyun Park and Jian Kang.

Group Inverse-Gamma Gamma (GIGG) Shrinkage

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

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

[3] [Armagan et al., 2013, Bai and Ghosh, 2019]

Group Inverse-Gamma Gamma (GIGG) Shrinkage

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

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

Result: [3] If $U \sim G(a, \eta)$ and $V \sim IG(b, \eta)$ are independent, then

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

[3] [Armagan et al., 2013, Bai and Ghosh, 2019]

Group Inverse-Gamma Gamma (GIGG) Shrinkage

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

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

Result: [3] If $U \sim G(a, \eta)$ and $V \sim IG(b, \eta)$ are independent, then

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

Formulation

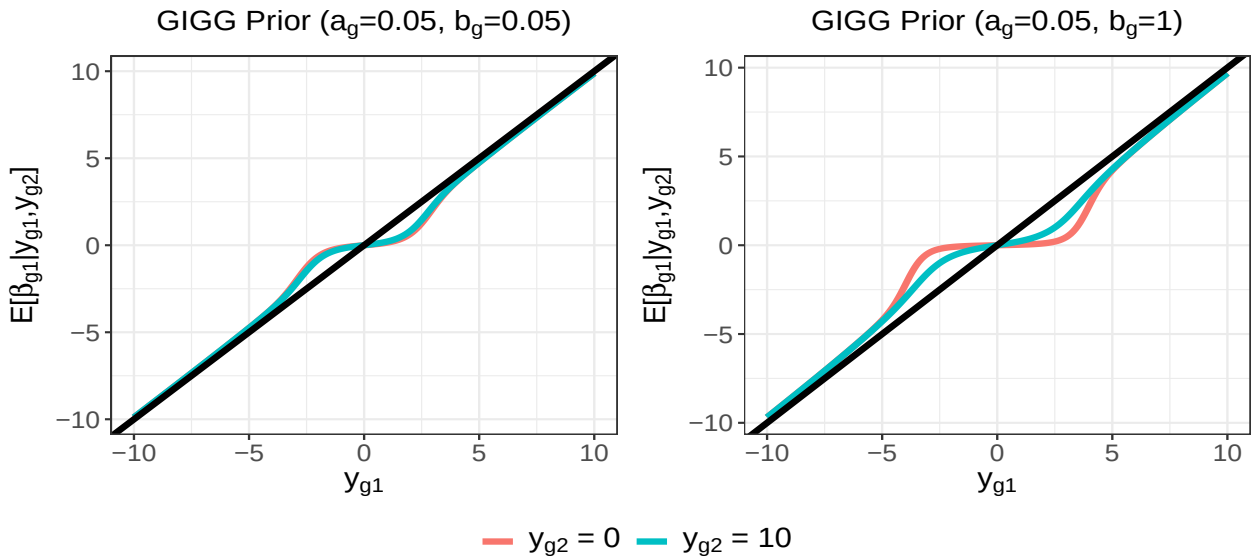
$$[\beta_{gj} \mid \tau^2, \gamma_g^2, \lambda_{gj}^2] \sim \mathcal{N}(0, \tau^2 \gamma_g^2 \lambda_{gj}^2), \quad \underbrace{\gamma_g^2 \sim G(a_g, 1)}_{\text{group}}, \quad \overbrace{\lambda_{gj}^2 \sim IG(b_g, 1)}^{\text{individual}}$$

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

[3] [Armagan et al., 2013, Bai and Ghosh, 2019]

Posterior Mean (GIGG Prior)

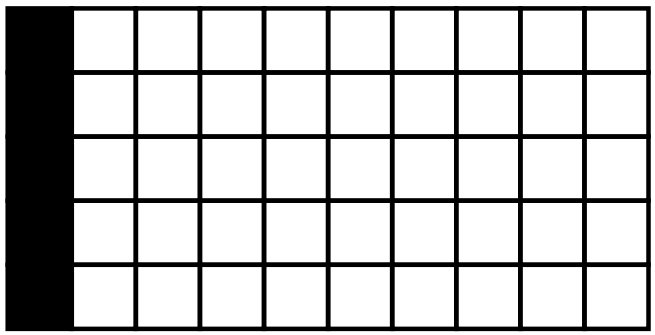
Illustrative Model: $[y_{g1}|\beta_{g1}] \sim N(\beta_{g1}, 1), \quad [y_{g2}|\beta_{g2}] \sim N(\beta_{g2}, 1)$



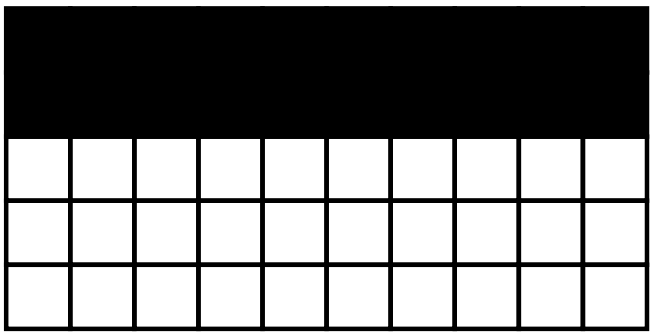
1. $a_g, b_g \approx 0 \Rightarrow$ shrinkage on β_{g1} hardly depends on $y_{g2} \Rightarrow$ individualistic shrinkage.
2. b_g moves away from zero $\Rightarrow$ shrinkage on β_{g1} more dependent on y_{g2}.
3. a_g: overall shrinkage, b_g: within-group shrinkage.

Simulations ($n = 500$, $p = 50$)

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 (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

*BGL-SS: Bayesian Group Lasso with Spike-and-Slab. BSGS-SS: Bayesian Sparse Group Selection with Spike-and-Slab

**Bolded entries indicate the top performers.

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 (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 performers.

▸ Skip to Shrinkage on Simplex

Illustrative Example from NHANES 2003-2004

Study Details

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

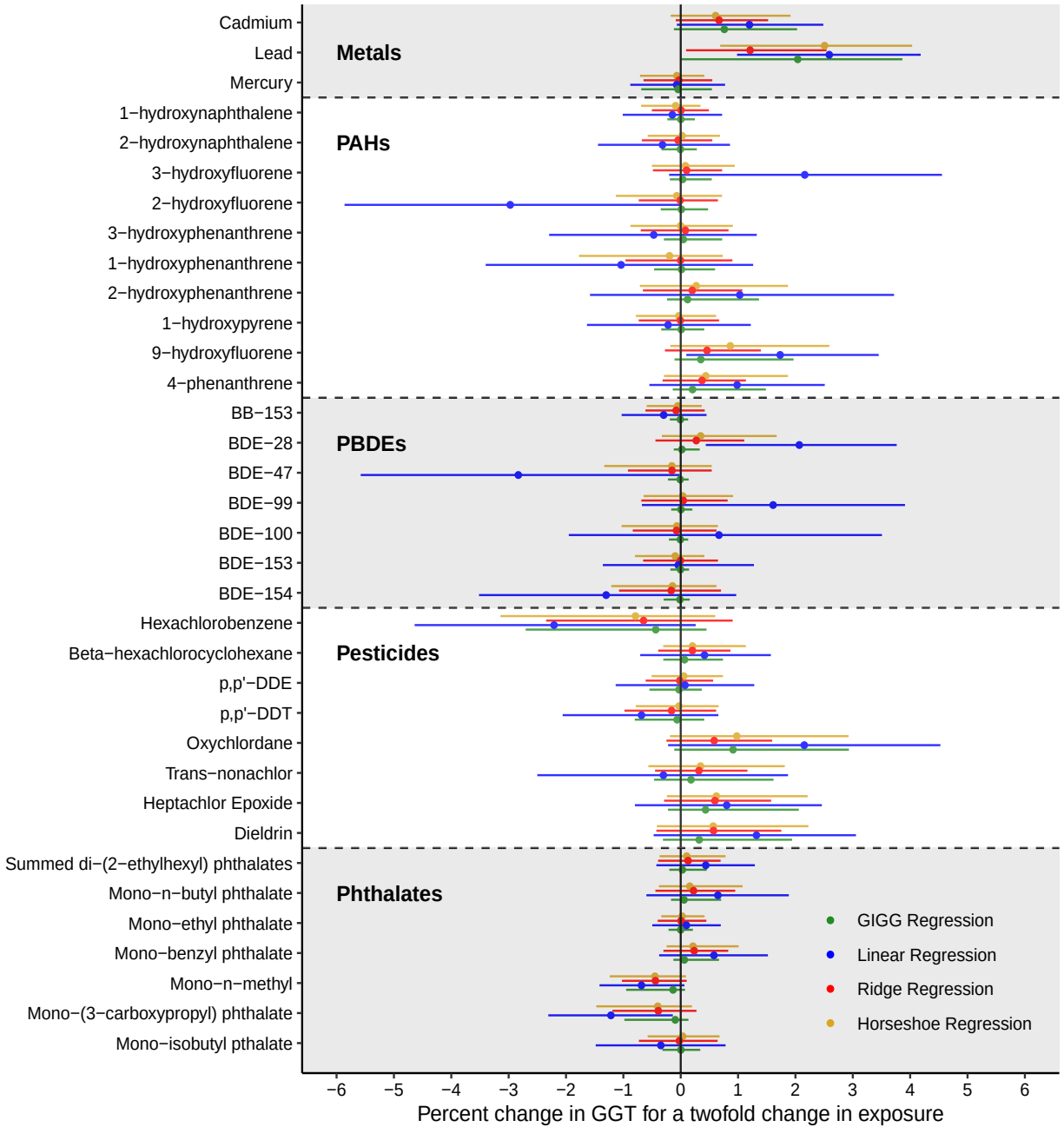
Exposure Classes

- 3 Metals[4] (*weak pairwise correlation and heterogeneous estimated effect sizes*)
- 7 Phthalates
- 8 Organochlorine Pesticides
- 7 Polybrominated Diphenyl Ethers (PBDEs)
- 10 Polycyclic Aromatic Hydrocarbons (PAHs)

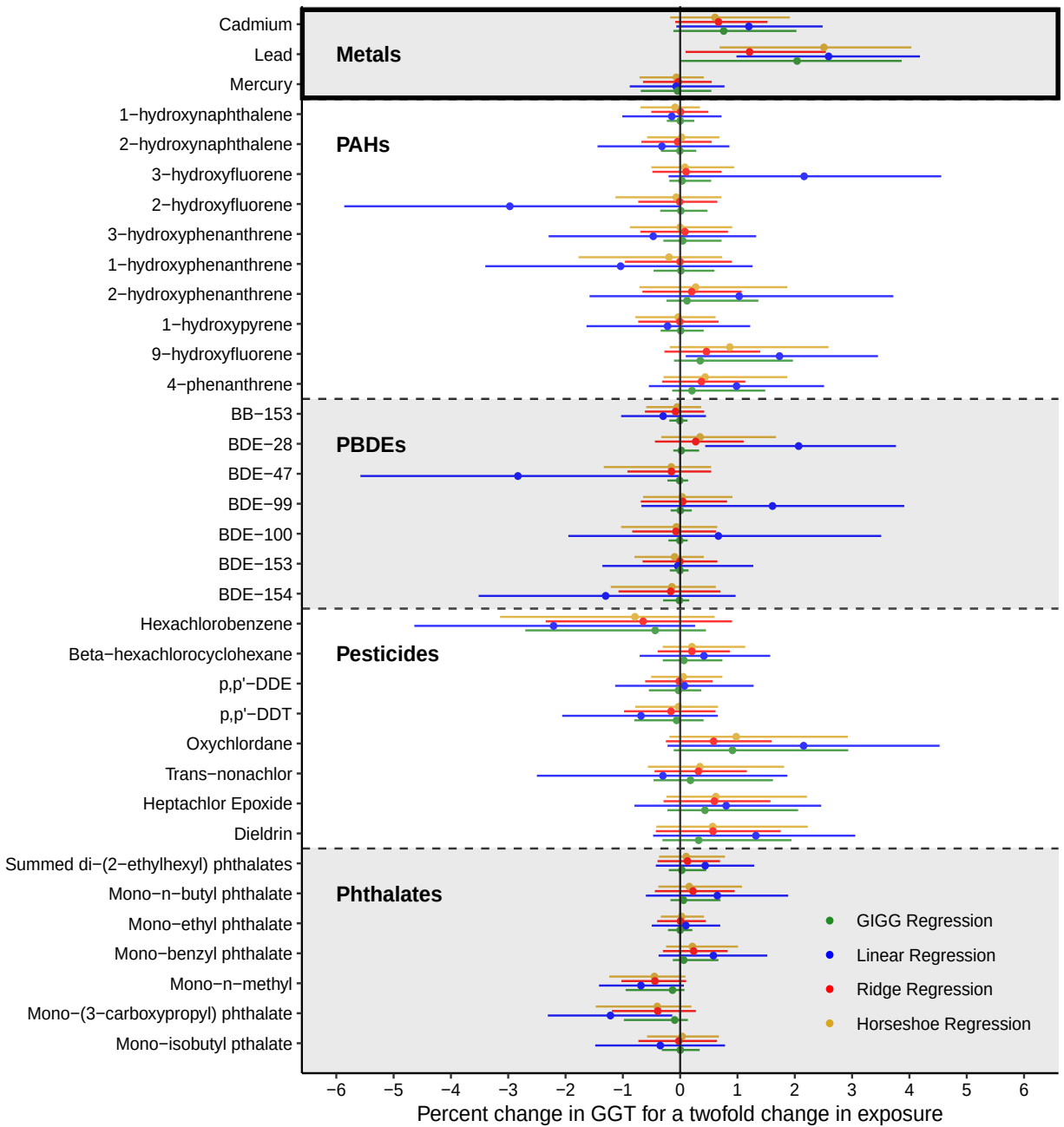
} *high pairwise correlation*

[4] cadmium, lead, and mercury

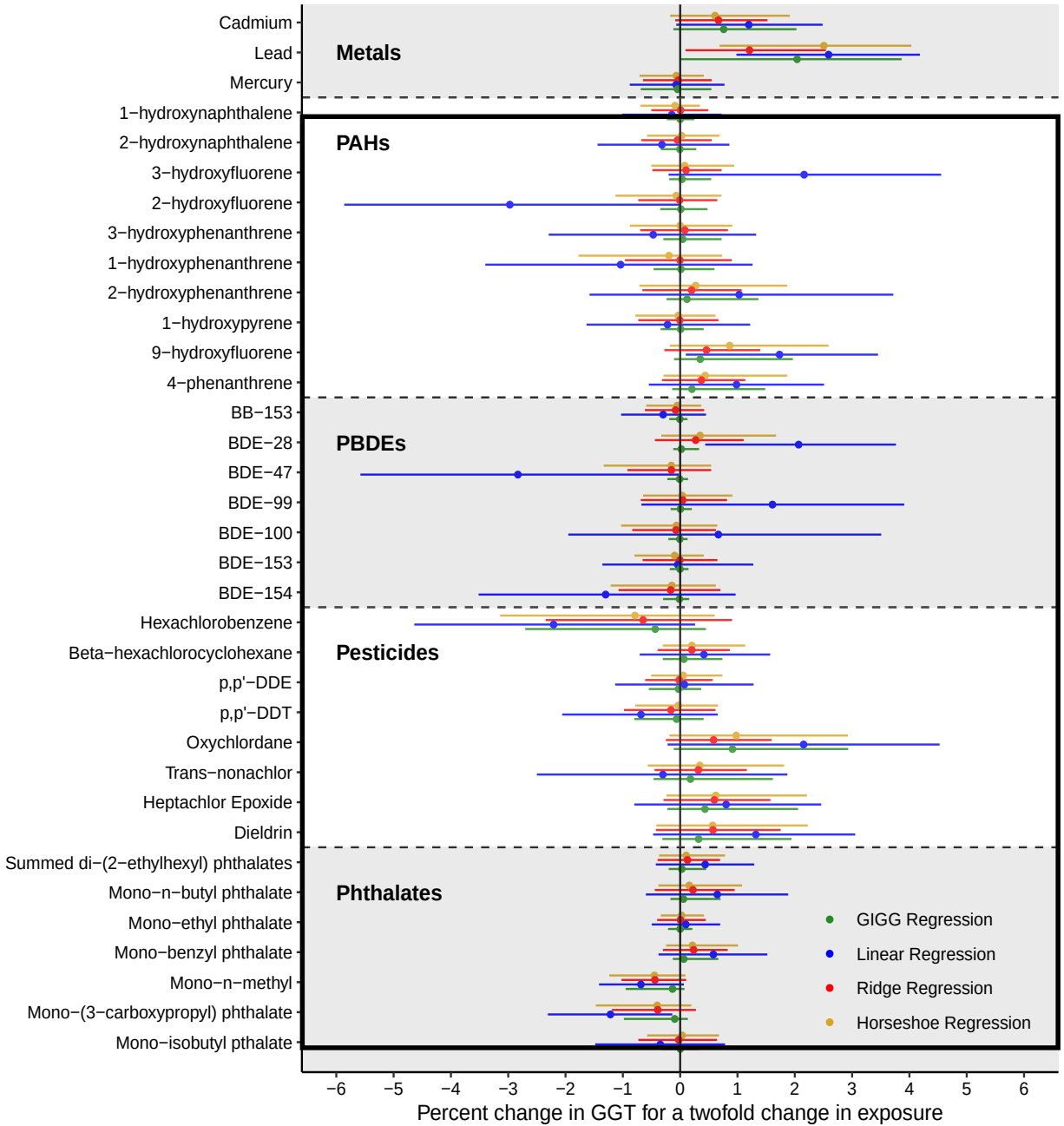
Illustrative Example from NHANES 2003-2004



Illustrative Example from NHANES 2003-2004



Illustrative Example from NHANES 2003-2004



Quasi-sparsity: Counts

Quasi-sparse Count Data

- Estimate an n-vector of Poisson means $\theta = (\theta_1, \ldots, \theta_n)^{\mathsf{T}}$, based on observations $y = (y_1, \ldots, y_n)^{\mathsf{T}}$, where $y_i \sim \text{Poi}(\theta_i)$ are independent.
- Quasi-sparse: Most θ_i's small but non-zero, few are large.
- Examples: Crime data, rare mutations, high-energy physics.
- ZI models have computational and identifiability issues for quasi-sparse data:

▸ Skip to Summary

Quasi-sparse Count Data

- Estimate an n-vector of Poisson means $\theta = (\theta_1, \ldots, \theta_n)^\mathsf{T}$, based on observations $y = (y_1, \ldots, y_n)^\mathsf{T}$, where $y_i \sim \text{Poi}(\theta_i)$ are independent.
- Quasi-sparse: Most θ_i's small but non-zero, few are large.
- Examples: Crime data, rare mutations, high-energy physics.
- ZI models have computational and identifiability issues for quasi-sparse data:
- Use shrinkage priors: in addition to a spike at zero and heavy tails, we need flexible thresholding for near-zero counts [Datta and Dunson, 2016, Biometrika].

▸ Skip to Summary

A bit of history!

- For $Y_i \sim \text{Poisson}(\theta_i)$, $E(\theta_i \mid Y_i)$ takes a simple form:

$$E(\theta_i \mid Y_i = y_i) = (y_i + 1)\frac{p_G(y_i + 1)}{p_G(y_i)}, \text{where } p_G(y) \text{ is marginal of } y.$$

[5] I. J. Good (1953): "The population frequencies of species and the estimation of population parameters". *Biometrika*

A bit of history!

- For $Y_i \sim \text{Poisson}(\theta_i)$, $E(\theta_i \mid Y_i)$ takes a simple form:

$$E(\theta_i \mid Y_i = y_i) = (y_i + 1)\frac{p_G(y_i + 1)}{p_G(y_i)}, \text{where } p_G(y) \text{ is marginal of } y.$$

- Empirical Bayes: $p_G(y_i) \approx \#\{Y_j = y_i\}/n$ (Robbins, 1956)

[5] I. J. Good (1953): "The population frequencies of species and the estimation of population parameters". *Biometrika*

A bit of history!

- For $Y_i \sim \text{Poisson}(\theta_i)$, $E(\theta_i \mid Y_i)$ takes a simple form:

$$E(\theta_i \mid Y_i = y_i) = (y_i + 1)\frac{p_G(y_i + 1)}{p_G(y_i)}, \text{where } p_G(y) \text{ is marginal of } y.$$

- Empirical Bayes: $p_G(y_i) \approx \#\{Y_j = y_i\}/n$ (Robbins, 1956)
- Used by Alan Turing and Jack Good [5] to crack German ciphers for the Enigma machine.

[5] I. J. Good (1953): "The population frequencies of species and the estimation of population parameters". *Biometrika*

A bit of history!

- For $Y_i \sim \text{Poisson}(\theta_i)$, $E(\theta_i \mid Y_i)$ takes a simple form:

$$E(\theta_i \mid Y_i = y_i) = (y_i + 1)\frac{p_G(y_i + 1)}{p_G(y_i)}, \text{where } p_G(y) \text{ is marginal of } y.$$

- Empirical Bayes: $p_G(y_i) \approx \#\{Y_j = y_i\}/n$ (Robbins, 1956)
- Used by Alan Turing and Jack Good [5] to crack German ciphers for the Enigma machine.
- Robbins' estimate performs well for low-dimensional and non-sparse data and can be optimized/smoothed for the sparse scenario [Brown, Greenshtein, & Ritov (2011), Koenker & Mizera (2014), Kiefer–Wolfowitz (1956) nonparametric MLE].

[5] I. J. Good (1953): "The population frequencies of species and the estimation of population parameters". *Biometrika*

Quasi-sparse Count data

Continuous Shrinkage for Poisson-Gamma hierarchical model:

$$y_i \sim \text{Poi}(\theta_i), \quad \theta_i \sim \text{Ga}(\alpha, \lambda_i^2\tau^2), \quad \lambda_i \sim p(\lambda_i^2), \quad \tau \sim p(\tau^2),$$

Marginalizing out θ_i and writing $\kappa_i = 1/(1+\lambda_i^2\tau^2)$:

$$p(y_i \mid \lambda_i, \tau) \propto \left(\frac{\lambda_i^2\tau^2}{1+\lambda_i^2\tau^2}\right)^{y_i} \left(\frac{1}{1+\lambda_i^2\tau^2}\right)^{\alpha},$$
$$p(y_i \mid \kappa_i) \propto (1-\kappa_i)^{y_i}\kappa_i^{\alpha} \Rightarrow [y_i \mid \kappa_i] \sim \text{NB}(\alpha, 1-\kappa_i).$$

The posterior distribution and mean of θ_i given y_i and κ_i are respectively

$$p(\theta_i \mid y_i, \kappa_i) \sim \text{Ga}(y_i+\alpha, 1-\kappa_i), \quad E(\theta_i \mid y_i, \kappa_i) = (1-\kappa_i)(y_i+\alpha).$$

Hence, the parameter κ_i can be interpreted as a random shrinkage factor pulling the posterior mean towards 0.

Flexible Shrinkage : GH prior

- The Gauss hypergeometric prior [Armero and Bayarri, 1994].

$$\text{GH}(\kappa_i \mid a, b, z, \gamma) = C\kappa_i^{a-1}(1-\kappa_i)^{b-1}(1+z\kappa_i)^{-\gamma} \text{ for } \kappa_i \in (0,1).$$

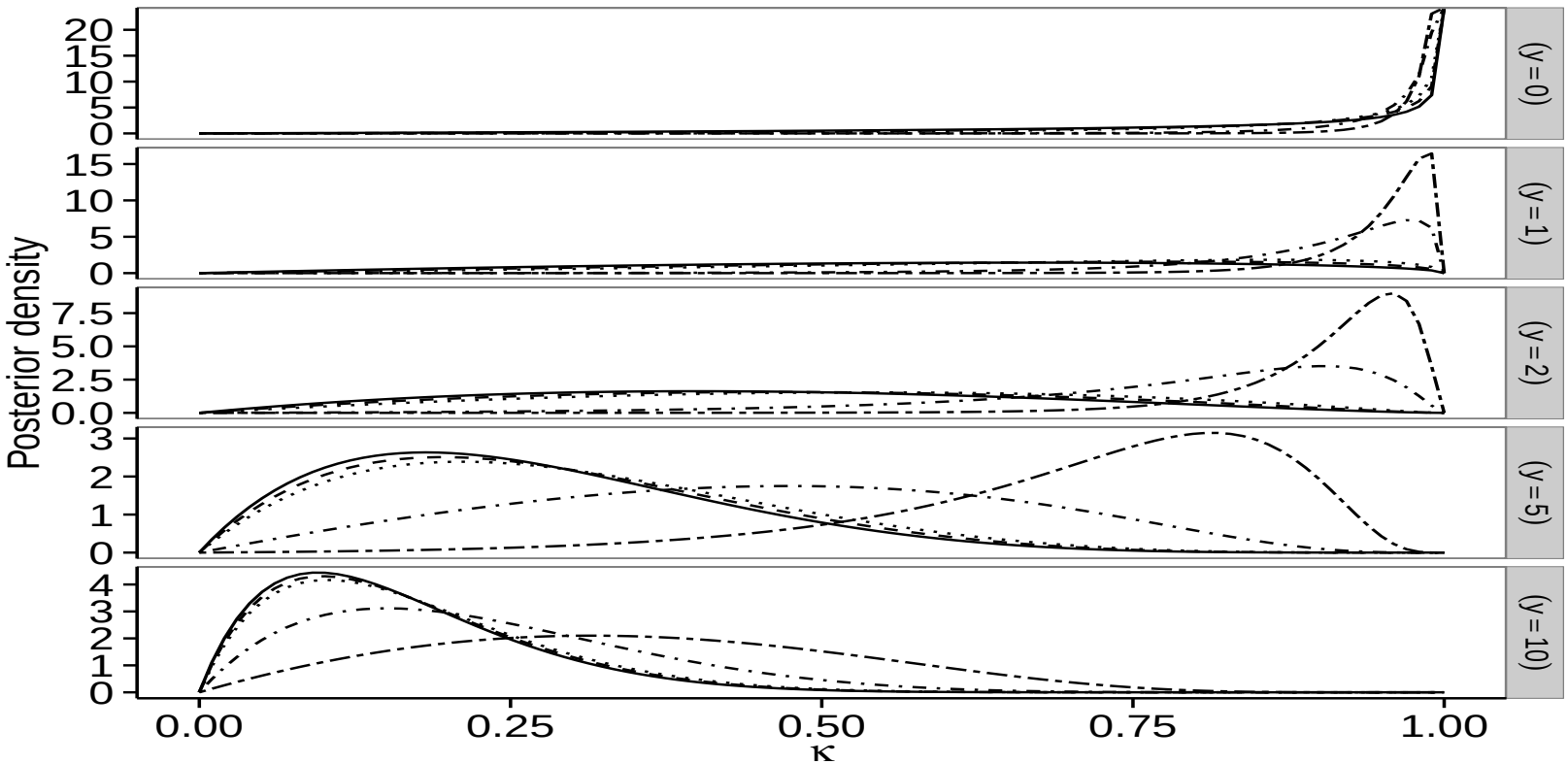


Figure 8: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), $\gamma = 10$ (two-dash).

Flexible Shrinkage : GH prior

- The Gauss hypergeometric prior [Armero and Bayarri, 1994].

$$\mathrm{GH}(\kappa_i \mid a, b, z, \gamma) = C\kappa_i^{a-1}(1-\kappa_i)^{b-1}(1+z\kappa_i)^{-\gamma} \text{ for } \kappa_i \in (0,1).$$

- γ enables flexible shrinkage by adapting to the quasi-sparsity.

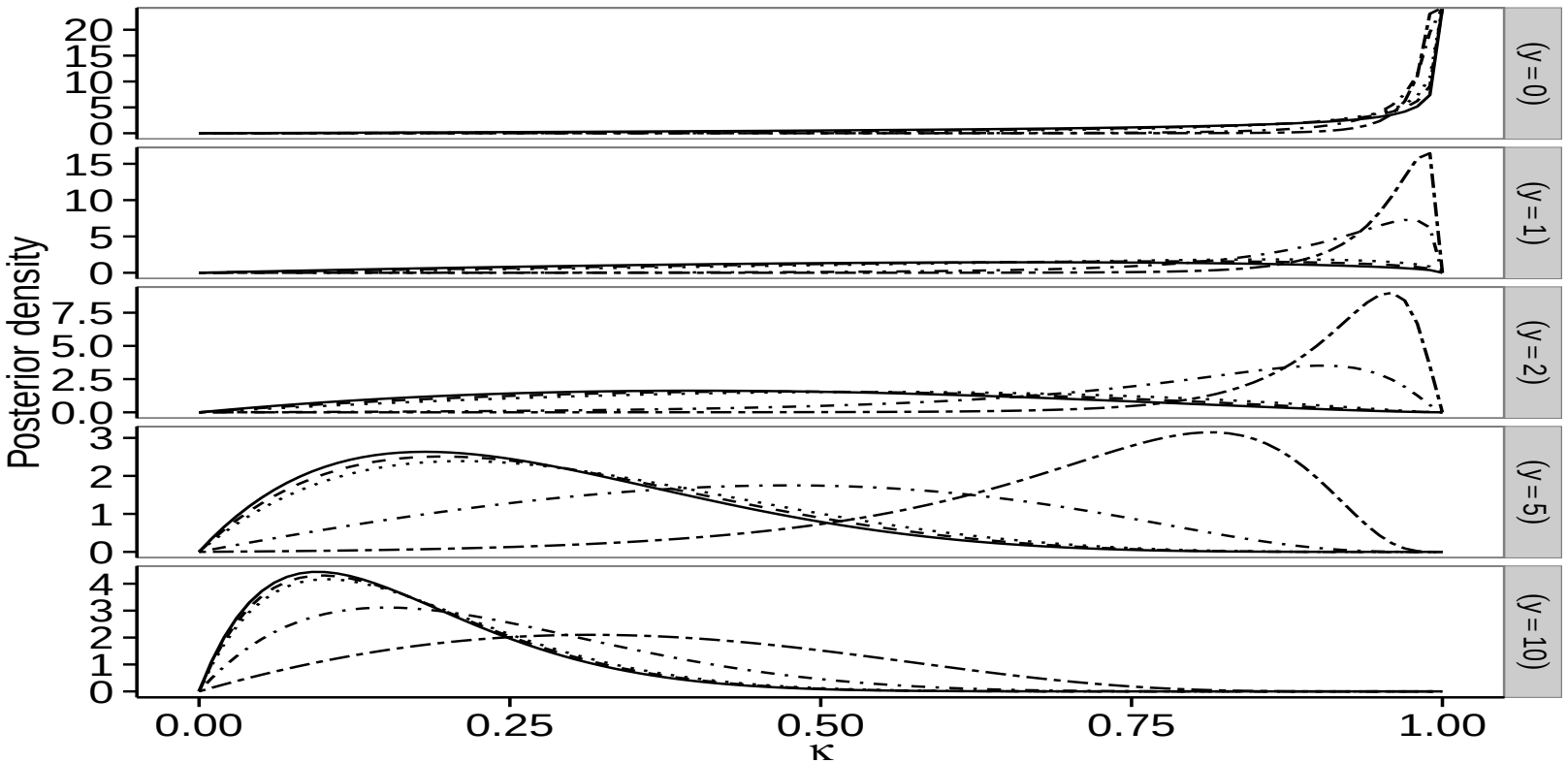


Figure 8: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), $\gamma = 10$ (two-dash).

Flexible Shrinkage : GH prior

- The Gauss hypergeometric prior [Armero and Bayarri, 1994].

$$\mathrm{GH}(\kappa_i \mid a, b, z, \gamma) = C\kappa_i^{a-1}(1-\kappa_i)^{b-1}(1+z\kappa_i)^{-\gamma} \text{ for } \kappa_i \in (0,1).$$

- γ enables flexible shrinkage by adapting to the quasi-sparsity.

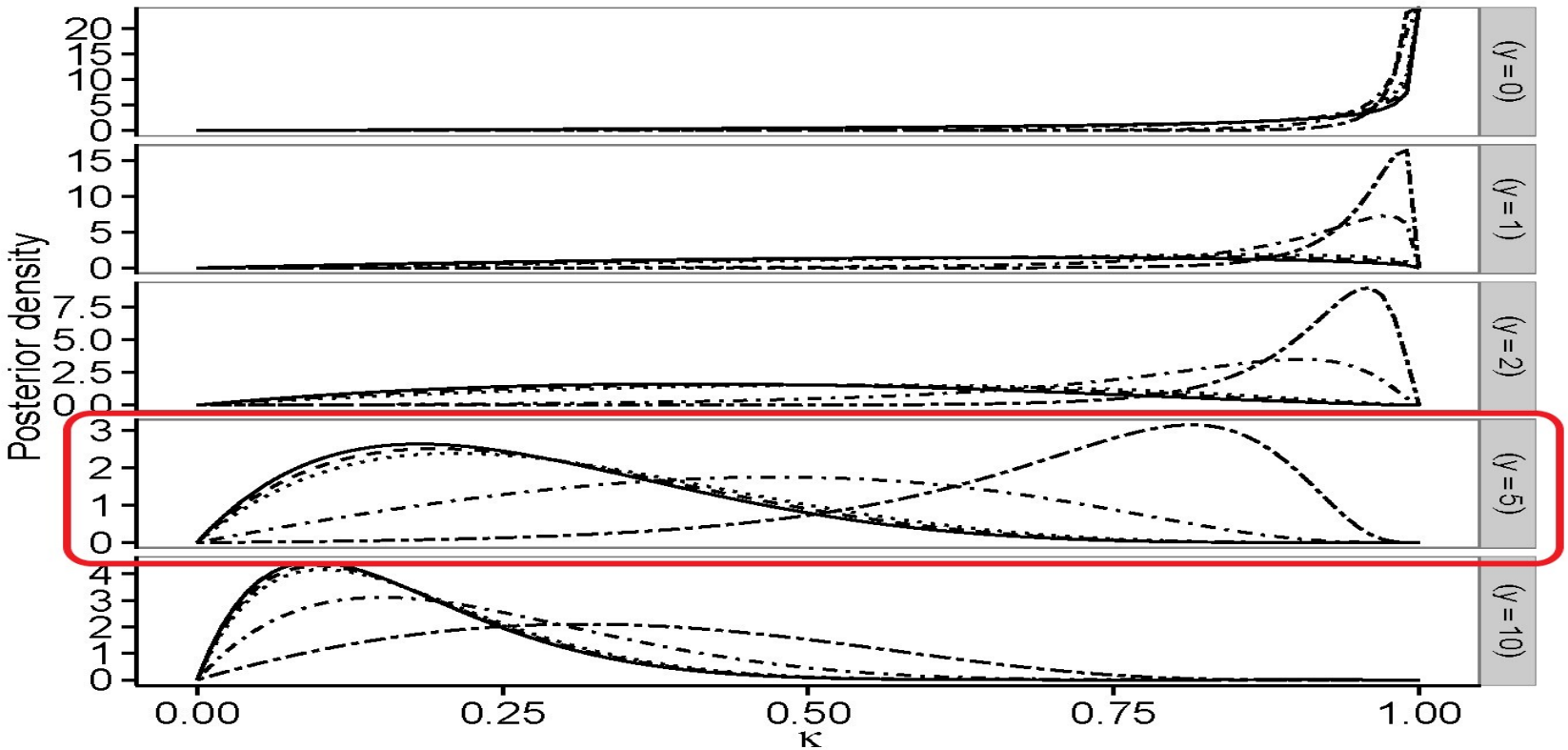


Figure 9: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), $\gamma = 10$ (two-dash).

Application: Rare Mutations

- Our goal is to identify the potential hotspots for rare mutations.
- Consider mutations with minor allele frequency $\leq 0.05\%$ Cirulli (2015) on a gene PIK3CA, which has been implicated for ovarian and cervical cancers.
- The distribution of mutational clusters for the GH method coincides with the true mutational clusters from the tumor portal.

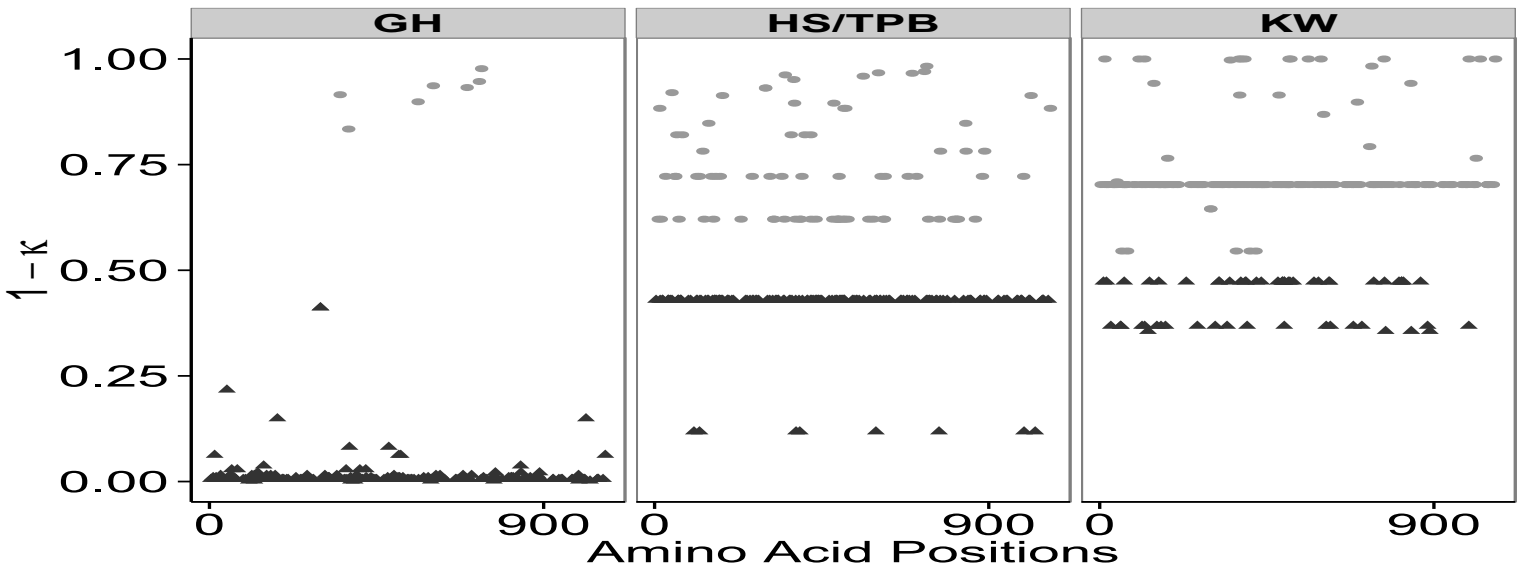


Figure 10: Comparison of Different Shrinkage Profiles

Shrinkage on Simplex

Shrinkage on simplex

Mixture Model: $\mathbf{y} \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_k f(y_i \mid \theta_k)$, K subgroups.

Compositional Data: $\mathbf{y} \sim \text{MN}(n; (\pi_1, \ldots, \pi_K))$, K species.

Network: $P(i \sim j) = \text{logit}^{-1}(\omega_{c_i, c_j}), \mathbf{c} \sim \text{Cat}(\pi_1, \ldots, \pi_K)$. K communities.

Here $\boldsymbol{\pi} \in \Delta_{K-1}$, i.e. $\sum_{j=1}^{K} \pi_j = 1$.

Goals:

1. Model sparsity in $\boldsymbol{\pi}$. [6]
2. Model a general dependence structure.

[6] *ongoing* work with Matt Heiner, Otso Ovaskainen, and David Dunson

Shrinkage on Simplex

- The Dirichlet distribution:

$$f(\pi_1, \pi_2, \ldots, \pi_K) \propto \pi_1^{\alpha_1 - 1} \cdot \pi_2^{\alpha_2 - 1} \cdots \pi_K^{\alpha_K - 1}, \ \alpha_i > 0, \sum \pi_j = 1.$$

- used routinely in categorical data analysis, also as a prior for mixture proportions and for the population distribution of latent variables.
- Popularity among applied modelers: (i) simple and easily interpretable structure, (ii) conjugacy to multinomial likelihoods facilitating computation.
- Main drawback: Symmetric Dir(α) can be inflexible for modeling sparse probabilities.
- Dir(α) density can be tuned to concentrate near 1-sparse vectors, but it is difficult to favor $\boldsymbol{\pi}$ being k-sparse.

Constructive Definition

- Goal: Introduce sparsity without over-parametrizing + retain conjugacy and neutrality of Dirichlet-Multinomial.
- **Stick-Breaking:** [Connor and Mosimann, 1969]

$$Z_k \sim f(\boldsymbol{\alpha}),\ Z_k \in (0,1),\ \boldsymbol{\alpha} \in \mathbb{R}^+$$

$$\pi_1 = Z_1,\ \pi_k = Z_k \prod_{l=1}^{k-1}(1 - Z_l),\ k = 1, \ldots, K-1,$$

$$\text{and } \pi_K = 1 - \sum_{k=1}^{K-1} \pi_k\,.$$

- $f(\cdot)$ is a density supported on the unit interval.
- $f = \mathrm{Be}\left(\frac{\alpha}{K},\ \alpha(1 - \frac{k}{K})\right) \Rightarrow \boldsymbol{\pi} \sim \mathrm{Dir}(\alpha)$.

Constructive Definition

- Goal: Introduce sparsity without over-parametrizing + retain conjugacy and neutrality of Dirichlet-Multinomial.
- **Stick-Breaking:** [Connor and Mosimann, 1969]

$$Z_k \sim f(\boldsymbol{\alpha}),\ Z_k \in (0,1),\ \boldsymbol{\alpha} \in \mathbb{R}^+$$

$$\pi_1 = Z_1,\ \pi_k = Z_k \prod_{l=1}^{k-1}(1 - Z_l),\ k = 1, \ldots, K-1,$$

$$\text{and } \pi_K = 1 - \sum_{k=1}^{K-1} \pi_k\,.$$

- $f(\cdot)$ is a density supported on the unit interval.
- $f = \mathrm{Be}\left(\frac{\alpha}{K},\ \alpha(1 - \frac{k}{K})\right) \Rightarrow \boldsymbol{\pi} \sim \mathrm{Dir}(\alpha)$.
- Idea: Use global-local shrinkage prior for $f(\cdot)$, e.g. Horseshoe or Gauss-hypergeometric prior.

Flexible Shrinkage : GH prior

- The Gauss hypergeometric prior [Armero and Bayarri, 1994].

$$\mathrm{GH}(\kappa_i \mid a, b, z, \gamma) = C\kappa_i^{a-1}(1-\kappa_i)^{b-1}(1+z\kappa_i)^{-\gamma} \text{ for } \kappa_i \in (0,1).$$

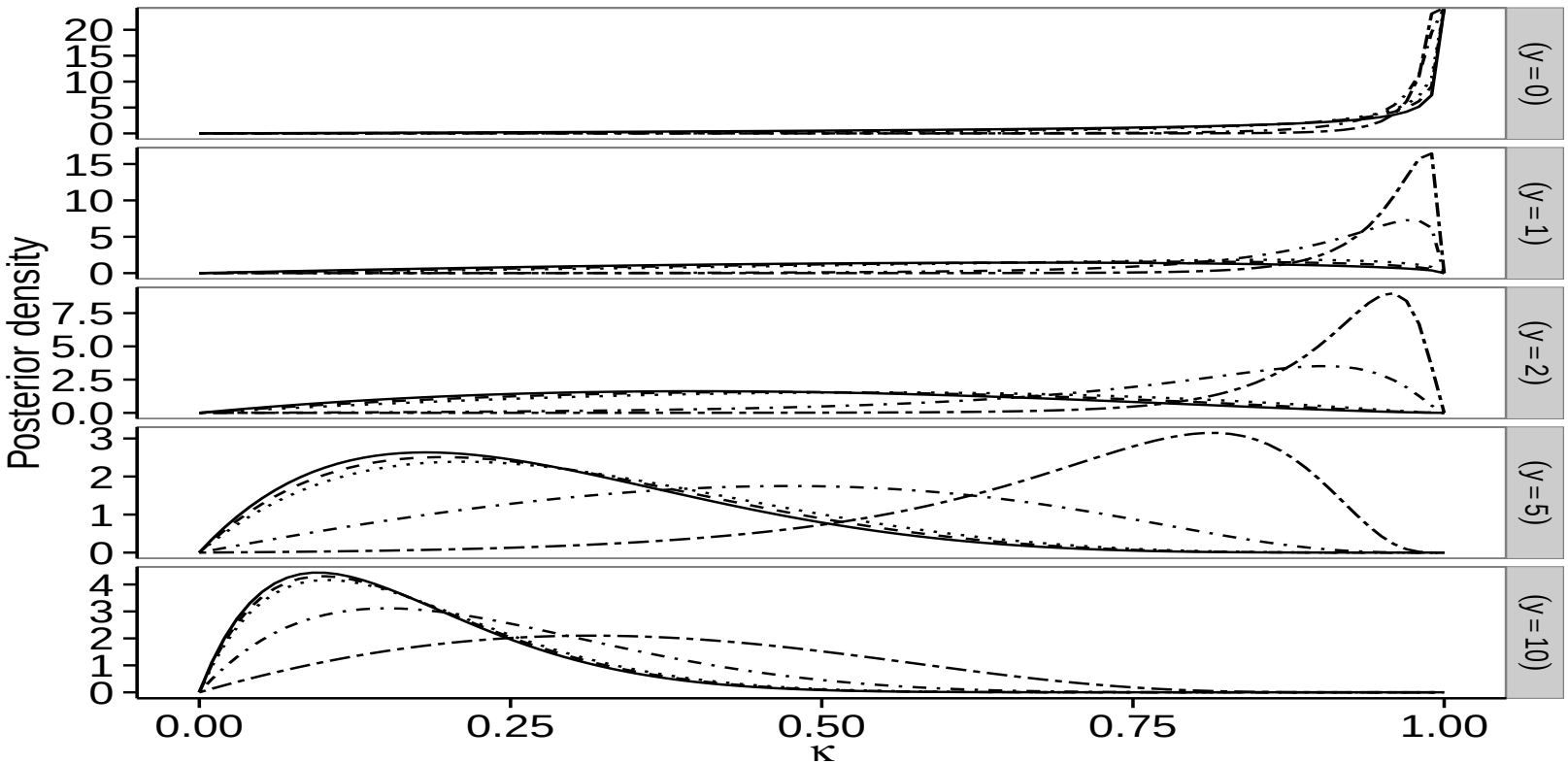


Figure 11: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), $\gamma = 10$ (two-dash).

Flexible Shrinkage : GH prior

- The Gauss hypergeometric prior [Armero and Bayarri, 1994].

$$\mathrm{GH}(\kappa_i \mid a, b, z, \gamma) = C\kappa_i^{a-1}(1-\kappa_i)^{b-1}(1+z\kappa_i)^{-\gamma} \text{ for } \kappa_i \in (0,1).$$

- γ enables flexible shrinkage by adapting to the quasi-sparsity.

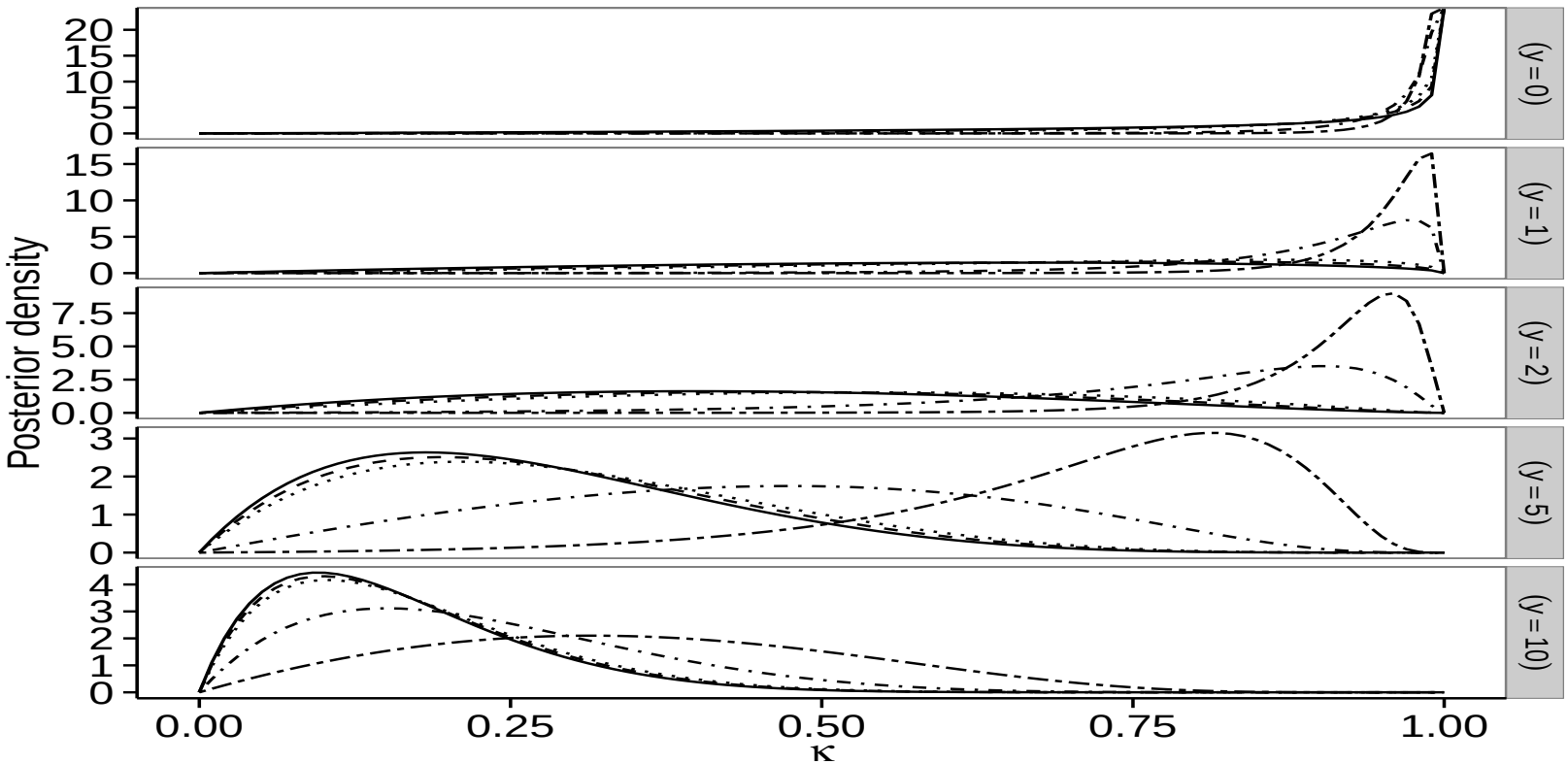
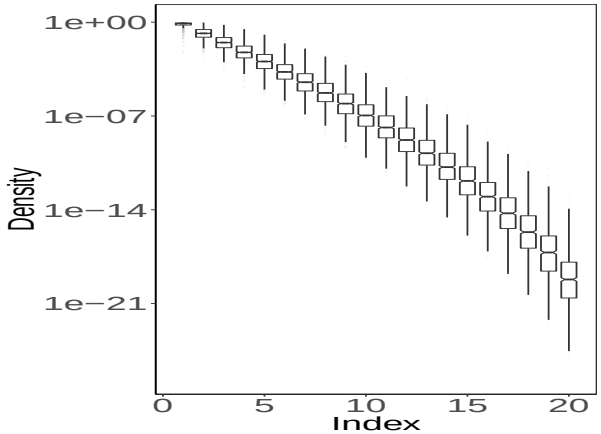


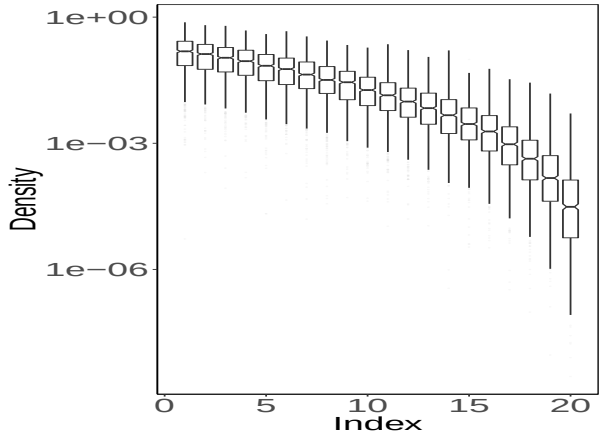
Figure 11: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), $\gamma = 10$ (two-dash).

Stochastic ordering in SGD

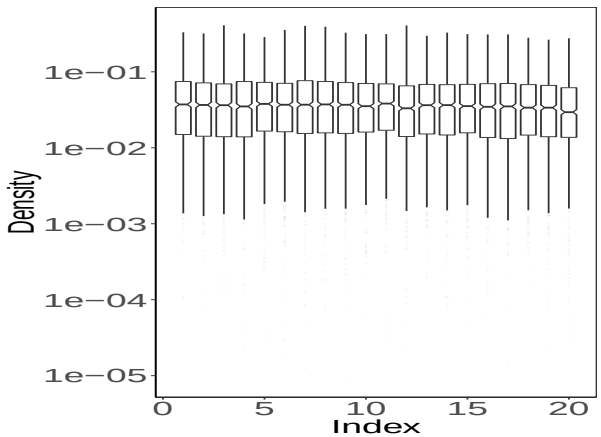
Using $Z_k \sim \text{GH}(\alpha/K,\ \alpha(1-k/K), \phi)$ induces an ordering and converges to a standard $\text{Dir}(\alpha)$ distribution as $\phi \to 1$.



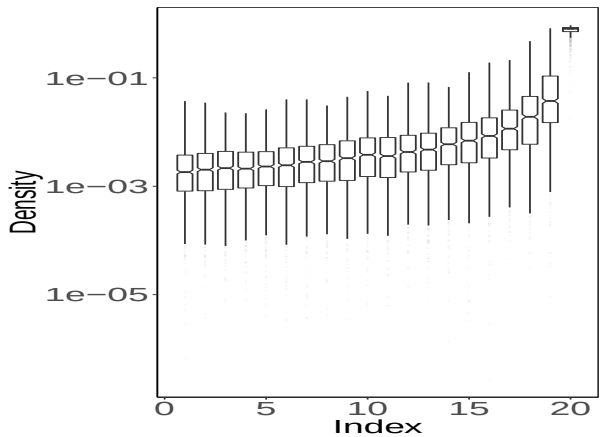
(a) ($\phi = 0.01$)



(b) ($\phi = 0.2$)



(c) ($\phi = 0.95$)



(d) ($\phi = 20.0$)

Extension: SSGD

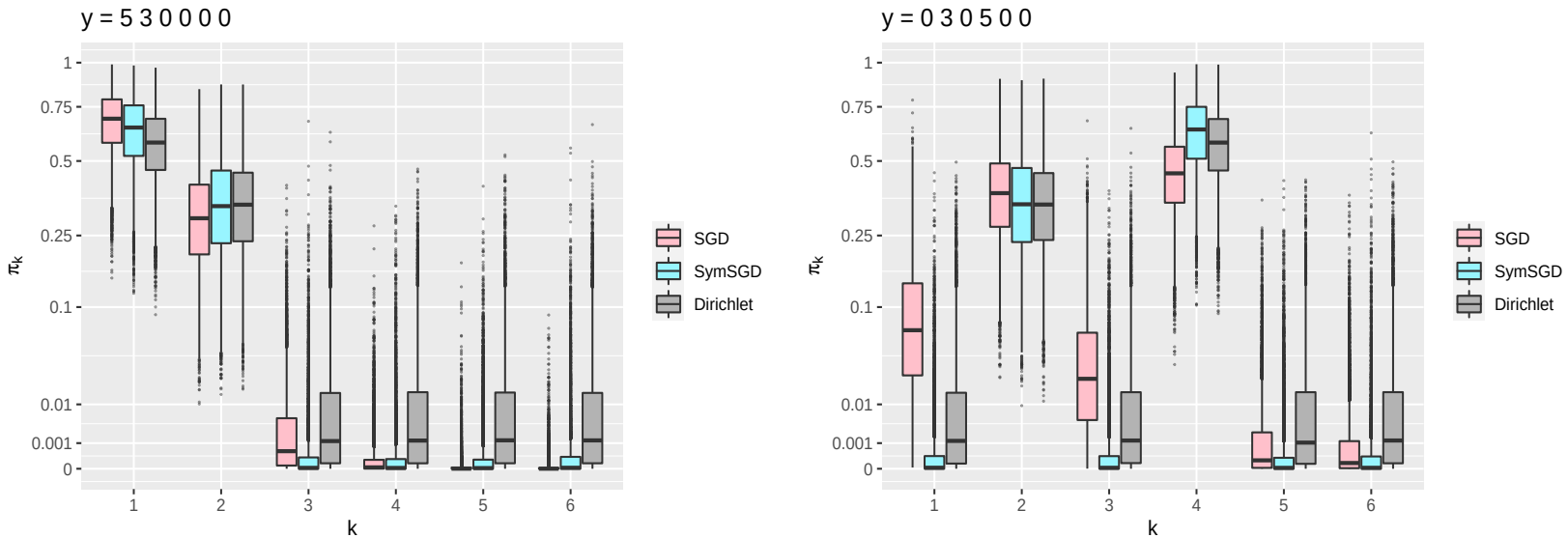
Consider placing a uniform prior on the indices of $\boldsymbol{\pi}$.

$$\begin{aligned} \boldsymbol{\pi}^* &\sim \mathcal{SGD}(\alpha, \phi)\,, \\ \boldsymbol{\sigma} &\sim \text{Unif}\,, \\ \boldsymbol{\pi} \mid \boldsymbol{\sigma}, \boldsymbol{\pi}^* &= \sigma(\boldsymbol{\pi}^*)\,, \end{aligned} \tag{3}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)$ is a permutation of the indices $\{1, \ldots, K\}$ and $\sigma(\cdot)$ reorders the elements of its argument according to $\boldsymbol{\sigma}$.

We call the marginal distribution of $\boldsymbol{\pi}$ under this model the symmetric sparse generalized Dirichlet distribution (SSGD).

Toy example: multinomial data



(a) Ordered

(b) Un-ordered

Figure 13: Box plots summarizing marginal posterior densities for proportions under three multinomial data scenarios (with $K = 6$) and three priors: SGD (pink), symmetric SGD (turquoise), and symmetric Dirichlet with shape parameter $1/K$ (gray).

Application: fungal biodiversity

Consider high-throughput sequencing data for dead wood inhabiting fungi by analyzing saw dust samples of Norwegian spruce from Ovaskainen et al. [2013].

> *The sequencing data reports the total sequencing depth for each sample, frequency for each of the 413 species and a few covariates, viz. the decay class, the part of the log where dust samples were collected from (basal or middle), and the genetic region used for molecular species identification (ITS1 or ITS2).*

Wood fungi

For studying co-occurrence using the first modeling approach, we consider only ITS2 samples and restrict to species that were present in at least 10 logs.

$$\mathbf{y}_i \mid \{\boldsymbol{\pi}_\ell\}, c_i \sim \text{Multinomial}(y_{i+}, \boldsymbol{\pi}_{c_i}), \; i = 1, \ldots, N \tag{4}$$

$$c_i \mid \boldsymbol{\lambda} \sim \text{Cat}(\lambda_1, \ldots, \lambda_L), \quad \text{where } \lambda_\ell = P(c_i = \ell), \tag{5}$$

$$\boldsymbol{\pi}_\ell \mid \phi_\ell \sim \mathcal{SSGD}(1, \phi_\ell), \; \ell = 1, \ldots, L, \tag{6}$$

$$\begin{aligned} \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L) &\sim \text{Dir}(1/L), \\ \phi_\ell &\sim \text{Unif}(1/K, 1), \; \ell = 1, \ldots, L. \end{aligned} \tag{7}$$

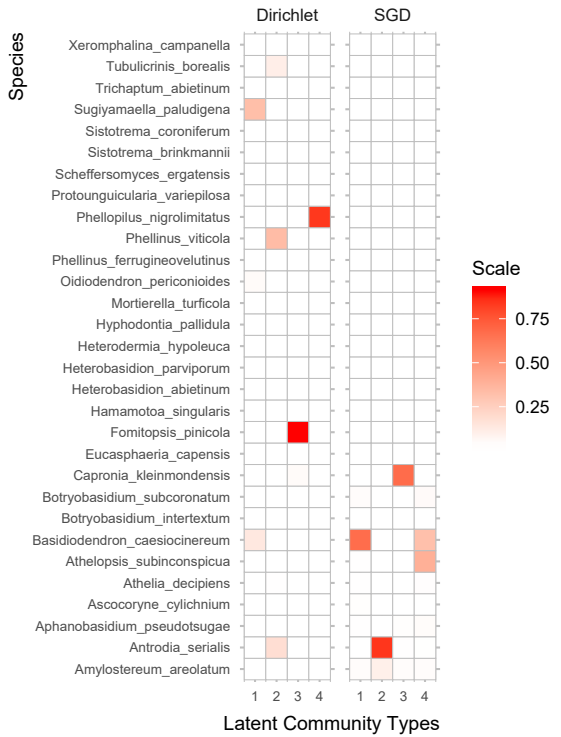


Figure 14: Posterior mean estimate for Π under Dirichlet and Sparse Generalized Dirichlet prior

General Dependence Structure

The Sparse Generalized Dirichlet appears to have a more general correlation structure compared to the Dirichlet prior, and explains some of the associations better.

(a) Dirichlet prior

(b) Sparse Generalized Dirichlet prior

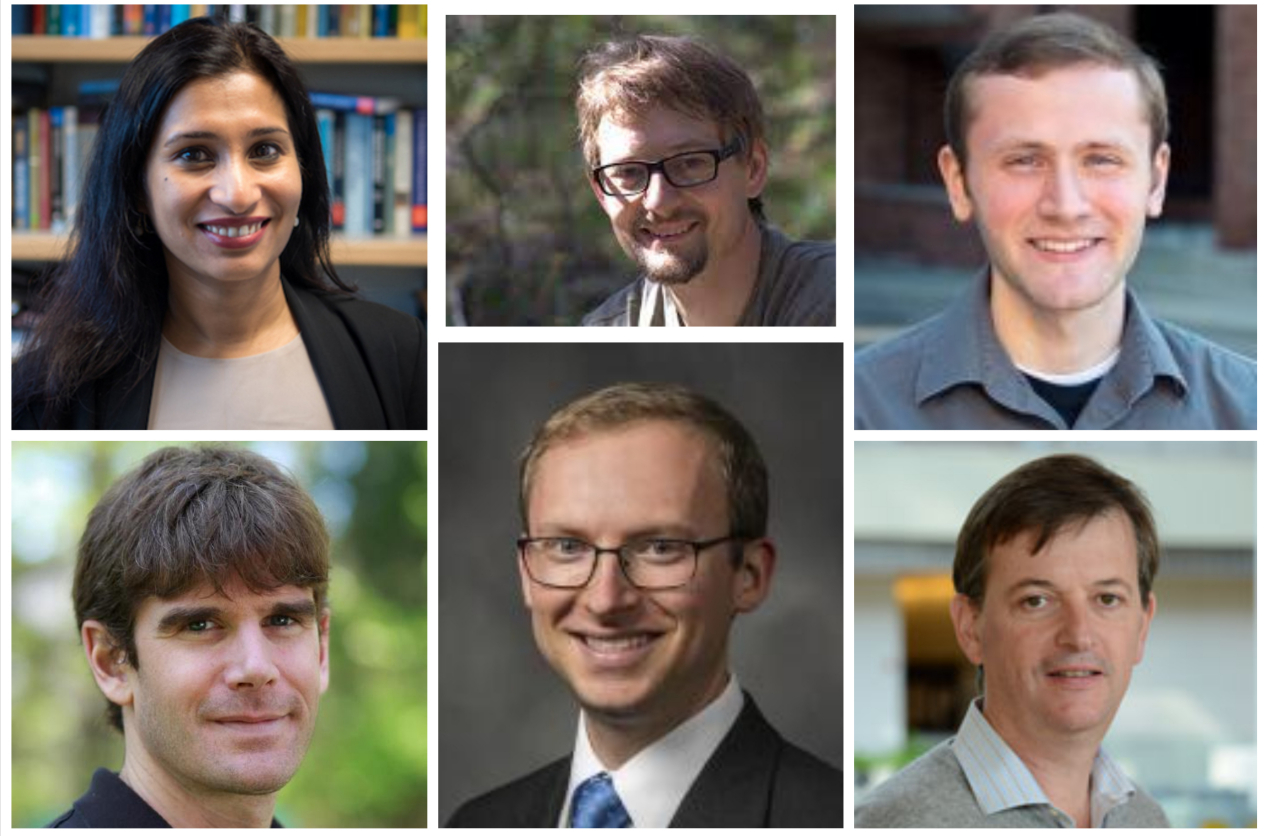
Summary and Scopes

- 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 + interpretation as non-convex penalty (horseshoe-like)
- Scopes:
  1. Selection for bi-level sparsity (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 (General global-Local)

- **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.
- 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!



Appendix

Two vs. One

- The computational advantage of one-group over two-groups is quite nuanced.
- Two-group priors do not need to visit all possible models, and one-group can often mix poorly.
- Continuous shrinkage priors allow many small signals $\Rightarrow$ more realistic and better performance in many settings.
- *"Sparsity can be construed in a weaker sense, where all of the entries in $\boldsymbol{\theta}$ are nonzero, yet most are small compared to a handful of large signals"* [Stephens and Balding (2009)]
- On the other hand, two-group priors work better in some situations like change-point detection [Guha and Datta, 2021].
- Empirical Bayes approaches are very promising [Martin and Tang, 2020].
- Normal scale mixture $\Rightarrow$ fast computation, leads to a non-convex penalty and can also work as a 'default' prior [Bhadra et al., 2016b].

Horseshoe Regularization

- Non-convex penalty : sparser model, need weaker coherence condition, low signal-to-noise ratio.
- Want to build a non-convex penalty with full probabilistic representation as the negative of the logarithm of a G-L prior.
- The prior $\pi(\theta)$ should have heavy-tails and spike at zero.
- Supports direct mode exploration (EM / Proximal Gradient) and MCMC for uncertainty quantification!

Horseshoe-like Priors i

- Recall: prior $p(\theta)$, induced penalty $-\log p(\theta)$.
- Horseshoe prior: $p(\theta)$ not analytically tractable - no closed form for penalty!

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

- Hindrance in learning via EM-type algorithms.

Horseshoe-like Priors ii

- Horseshoe prior admits tight upper and lower bounds, normalize them:

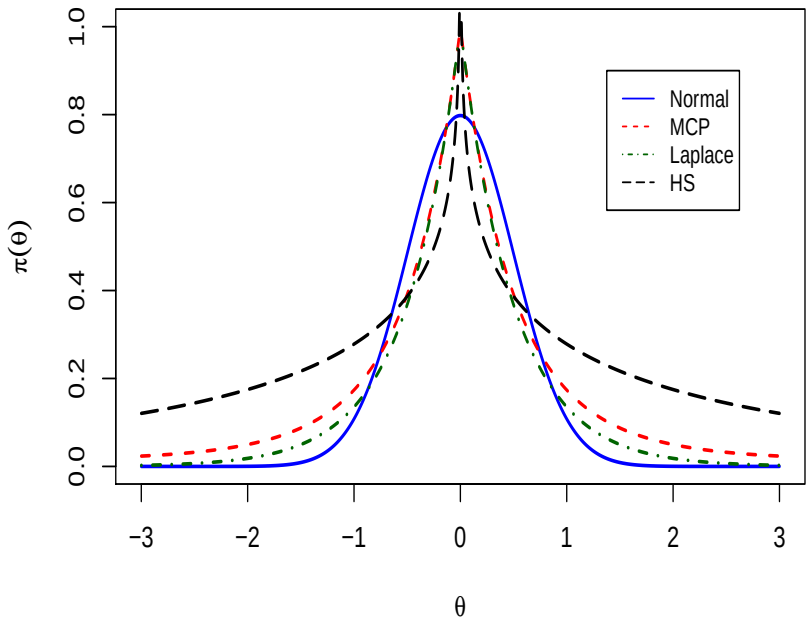
$$\frac{1}{\tau(2\pi)^{3/2}} \log\left(1+\frac{4\tau^2}{\theta_i^2}\right) < p_{HS}(\theta_i \mid \tau) < \frac{2}{\tau(2\pi)^{3/2}} \log\left(1+\frac{2\tau^2}{\theta_i^2}\right).$$

- 'Horseshoe-like' prior on θ_i:

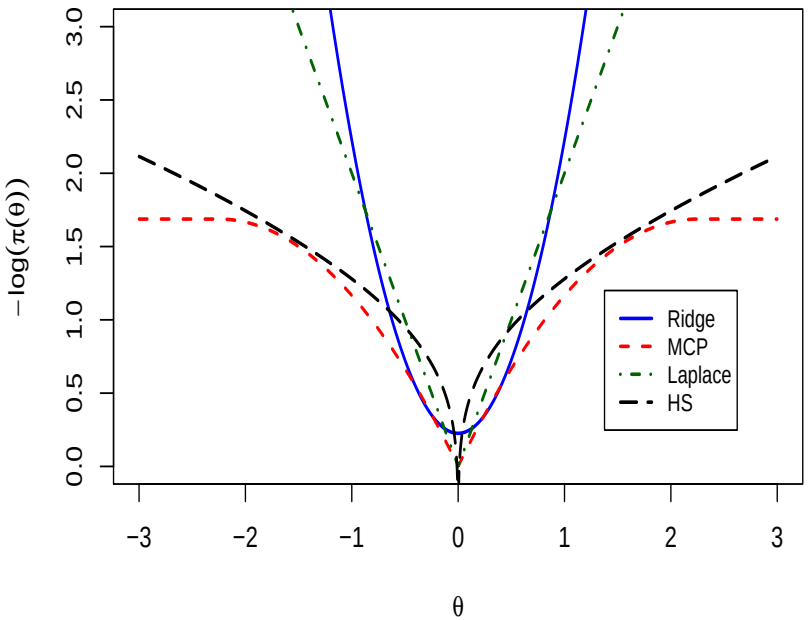
$$p_{\widetilde{HS}}(\theta_i \mid a) = \frac{1}{2\pi a^{1/2}} \log\left(1+\frac{a}{\theta_i^2}\right),$$

- $a = 2\tau^2$ and $a = 4\tau^2$ in (50) recovers the bounds in (**??**).

Horseshoe Penalty



(a) Prior densities



(b) Induced Penalty

- Horseshoe penalty is more aggressive near zero compared to the convex penalties, encouraging sparsity.

Horseshoe-like prior

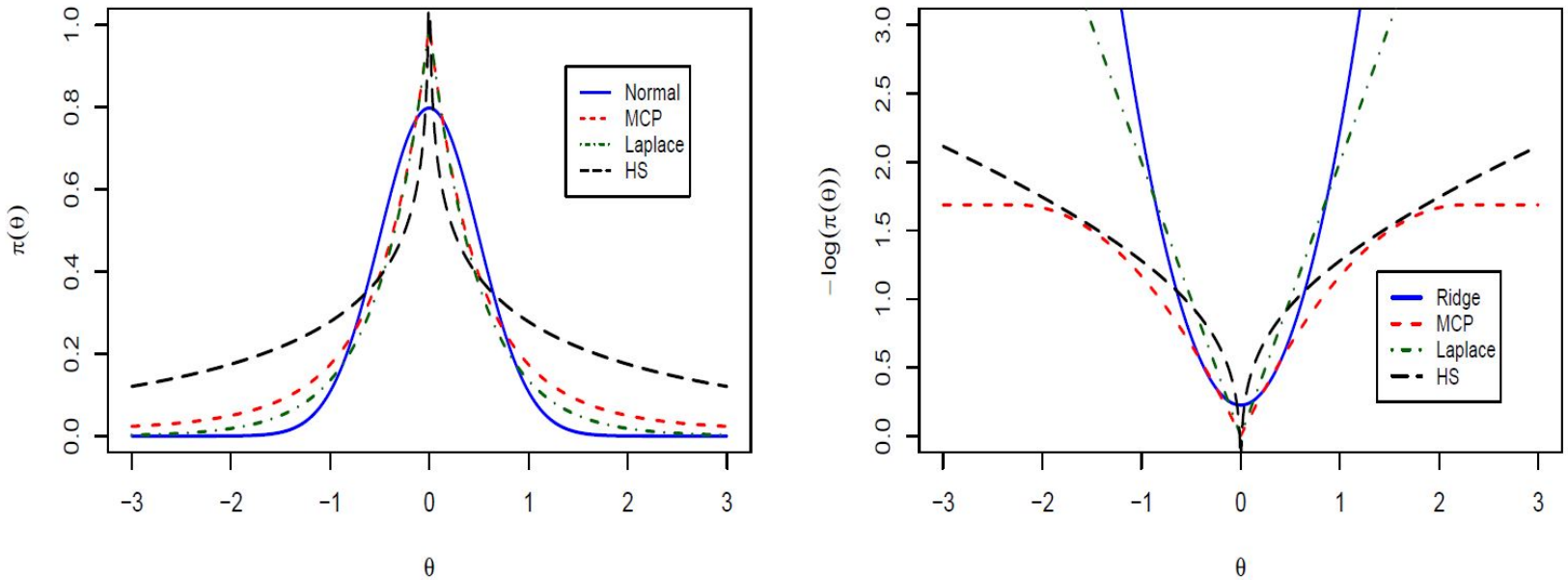


Figure 16: (a) Prior density (b) Induced Penalty

- Horseshoe penalty is more aggressive near zero compared to the convex penalties, encouraging sparsity.
- We still need scale mixture representation for efficient computation !

Scale Mixture Representation!

- Frullani's identity

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

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

$$\frac{1}{2\pi a^{1/2}} \log\left(1 + \frac{a}{\theta_i^2}\right) = \int_0^\infty \exp\left(-\frac{u_i \theta_i^2}{a}\right) \frac{(1 - e^{-u_i})}{2\pi a^{1/2} u_i} du_i$$

- Normal scale mixture:

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

Resources for Horseshoe Prior

The Bayes Oracle

- *Bayes oracle*: Risk for BO = Lower bound of $(1/m)$ times the risk for any multiple testing procedure under the two-groups model.
- Makes use of the unknown parameters of the mixture, ψ^2 and p, not attainable in finite samples.
- Reparametrize $u = \psi^2$ and $v = uf^2$, Bayes oracle threshold becomes:

$$C^2 = \left(1 + \frac{1}{u}\right)\left(\log v + \log(1 + \frac{1}{u})\right)$$

- If $C = 0$ then both the errors are zero and for $C = \infty$, the inference is essentially no better than tossing a coin.

Asymptotic framework [Bogdan et al., 2011]

Assumption

Proportion of signals $= p = p_m$, *magnitude of signals* $= \psi = \psi_m$ *(non-null variance).*

$$p_m \to 0 \text{ (sparsity)} \; ; \; \psi_m^2 \to \infty \text{ (signal strength)}$$
$$\frac{2\log(p_m^{-1})}{\psi_m^2} \to C \in (0, \infty) \text{ as } m \to \infty$$

- Example: $p_m = m^{-\beta}, \beta \in (0,1)$ and $\psi_m^2 = 2\log m$ would satisfy the above.
- Under Assumption 1, the Bayes risk for Bayes Oracle is

$$R_{opt} = m\left\{(1-p)t_1^{BO} + pt_2^{BO}\right\} = mp\{2\Phi(\sqrt{C}) - 1\}(1 + o_m) \quad (8)$$

- If $C = 0$ then both the errors are zero and for $C = \infty$, the inference is essentially no better than tossing a coin.

▸ Optimality of Horseshoe

Asymptotic Optimality

Theorem

Prob. of type I error: $t_1 = \left\{2\tau^2/\sqrt{\ln(1/\tau)}\right\}(1+o(1))$

Theorem

Prob. of type II error: $t_2 \leq (2\Phi(\sqrt{\frac{2}{\eta(1-\delta)}}\sqrt{C}) - 1)(1+o(1))$
($\eta, \delta \in (0,1)$ fixed const)

- Bayes Risk of HS+:
  $R_{HS+} = m\left\{p(2\Phi(\sqrt{\frac{2}{\eta(1-\delta)}}\sqrt{C}) - 1)\right\}(1+o(1))$
- **Bayes Oracle:** Lower bound of Bayes Risk. (under known parameter values).
- Bayes Risk of the Oracle $R_{BO} = m\left\{p(2\Phi(\sqrt{C}) - 1)\right\}(1+o(1))$.
- HS+ decision rule will attain the Bayes oracle up to a multiplicative constant (≈ 1 in nearly black case). ▸ Recall: Bayes Oracle

Gibbs sampler

- The original hierarchy for horseshoe does not allow Gibbs sampling: alternatives like slice sampling were used.
- Define $\eta_j = 1/\lambda_j^2$ and $\mu_j = \beta_j/(\sigma\tau)$, then,

$$p(\eta_j \mid \tau, \sigma, \mu_j) \propto \exp\left\{-\frac{\mu_j^2}{2}\eta_j\right\} \frac{1}{1+\eta_j}.$$

- Slice sampling:
  1. Sample $u_j \mid \eta_j \sim \mathcal{U}(0, \frac{1}{1+\eta_j})$,
  2. Sample $\eta_j \mid \mu_j, u_j \sim \mathrm{Exp}(2/\mu_j^2)$ truncated to $[0, \frac{1-u_j}{u_j}]$.
- Makalic and Schmidt [2016] observed that: if $\tau^2 \mid \xi \sim \text{InvGamma}(1/2, 1/\xi)$ and $\xi \sim \text{InvGamma}(1/2,1)$ then marginally $\tau \sim \mathcal{C}^+(0,1)$.
- This leads to an efficient Gibbs sampler!

Gibbs Sampler Makalic and Schmidt [2016]

The above hierarchy makes Gibbs sampling from the posterior distribution straightforward. The conditional posterior distribution of the regression coefficients $\beta \in \mathbb{R}^p$ [8] is

$$\beta|\cdot \sim \mathcal{N}_p(\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \sigma^2\mathbf{A}^{-1}), \quad \mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{\Lambda}_*^{-1}), \quad \mathbf{\Lambda}_* = \tau^2\mathbf{\Lambda}, \tag{9}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_p^2)$. The conditional posterior distribution of σ^2 is an inverse-gamma distribution given by

$$\sigma^2|\cdot \sim \mathcal{IG}\left((n+p)/2, (\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta)/2 + \beta^{\mathrm{T}}\mathbf{\Lambda}_*^{-1}\beta/2\right). \tag{10}$$

The conditional posterior distributions for the local and global hypervariances are also of inverse-gamma type

$$\begin{aligned} \lambda_j^2|\cdot &\sim \mathcal{IG}\left(1, \frac{1}{\nu_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \quad (j = 1, 2, \ldots, p), \\ \tau^2|\cdot &\sim \mathcal{IG}\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\lambda_j^2}\right). \end{aligned} \tag{11}$$

Finally, the conditional posterior distributions for the auxiliary variables are:

$$\begin{aligned} \nu_j|\cdot &\sim \mathcal{IG}\left(1, 1 + \frac{1}{\lambda_j^2}\right), \quad (j = 1, 2, \ldots, p), \\ \xi|\cdot &\sim \mathcal{IG}\left(1, 1 + \frac{1}{\tau^2}\right). \end{aligned}$$

Figure 17: Makalic and Schimdt sampler

Bhattacharya et al. [2016] trick

- Conditional posterior of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$, τ and σ.

$$\boldsymbol{\beta} \mid y, \lambda, \tau, \sigma \sim \mathcal{N}(A^{-1}X^T y, \sigma^2 A^{-1}),$$
$$A = (X^T X + \Lambda_*^{-1}), \Lambda_* = \tau^2 \mathrm{diag}(\lambda_1^2, \cdots, \lambda_p^2).$$

- Generic problem: want samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \underbrace{(\Phi^T \Phi + D^{-1})^{-1}}_{O(p^3)}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\Phi^T \boldsymbol{\alpha}$.
- Alternative algorithm: complexity $O(n^2 p)$ when D is diagonal, beneficial when $p \gg n$.

Algorithm 1 Proposed algorithm

(i) Sample $u \sim \mathrm{N}(0, D)$ and $\delta \sim \mathrm{N}(0, \mathrm{I}_n)$ independently.
(ii) Set $v = \Phi u + \delta$.
(iii) Solve $(\Phi D \Phi^{\mathrm{T}} + \mathrm{I}_n) w = (\alpha - v)$.
(iv) Set $\theta = u + D\Phi^{\mathrm{T}} w$.

Figure 18: Bhattacharya et al. [2016] algorithm

- See Johndrow et al. [2020] for a Markov approximation for even faster algorithm.

Implementation

Table 1: 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]

GIGG Supplementary

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.

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.

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.

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.

Theory for GIG

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.