Bayesian Sparse Signal Recovery: Gaussian Models and Beyond

Jyotishka Datta[1]

August 4, 2020

University of Arkansas, Fayetteville

Outline

Bayesian Sparse Signal Recovery: Gaussian Models and Beyond

Part I: Background: global-local priors

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

Part II: Discrete data

1. Quasi-sparse count
2. Gauss hypergeometric Prior.

Part III: Shrinkage on Simplex

1. Sparse Generalized Dirichlet
2. Contraction properties
3. Numerical Results
4. Theory
5. Future directions

Global-Local Shrinkage: A Brief Overview

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

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

Today

Poisson means $(Y_i \mid \theta_i) \overset{\text{ind}}{\sim} \text{Poi}(\theta_i), i = 1, \ldots, n,$

Quasi-sparsity: $\boldsymbol{\theta} \in \Theta[p_n, \epsilon_n] \equiv \{\boldsymbol{\theta} : \#(|\theta_i| > \epsilon_n) \leq p_n\}, p_n/n \to 0$

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

Goals:

1. Recovery: provide estimator $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\pi}}$.
2. Multiple Testing: Test whether each θ_i (or π_i) is zero or non-zero.
3. Variable selection / Prediction.
4. Model sparsity in $\boldsymbol{\pi}$.

One-group model

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

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

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

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

One-group model

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

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

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

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?

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. [2017a]
Horseshoe-like	Bhadra et al. [2017b]
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]

Theory for general G-L prior

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

- Datta and Ghosh [2013], Ghosh et al. [2016]: G-L priors asymptotically Bayes optimal under sparsity (ABOS).
- Ghosh and Chakrabarti [2017], van der Pas et al. [2016a,a,c, 2017]: posterior concentration at near-minimax rate.
- Need: 'heavy-tailed prior with sufficient mass near zero'
- Up to $O(1)$ can be relaxed: G-L priors can be exactly minimax and ABOS when τ is treated as a tuning parameter [Ghosh and Chakrabarti, 2017, Bai and Ghosh, 2017].

Discrete Data

Quasi-sparse Count Data

- We can extend G-L priors for non-Gaussian data, e.g. quasi-sparse count data?
- Quasi-sparse: $\theta \in \Theta[p_n, \epsilon_n] \equiv \{\theta : \#(|\theta_i| > \epsilon_n) \leq p_n\}, p_n/n \to 0$. 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 problems for quasi-sparse data.

▸ Skip to Shrinkage on Simplex

Quasi-sparse Count Data

- We can extend G-L priors for non-Gaussian data, e.g. quasi-sparse count data?
- Quasi-sparse: $\theta \in \Theta[p_n, \epsilon_n] \equiv \{\theta : \#(|\theta_i| > \epsilon_n) \leq p_n\}, p_n/n \to 0$. 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 problems 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, Bka].

▸ Skip to Shrinkage on Simplex

Flexible Shrinkage : GH prior

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

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

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

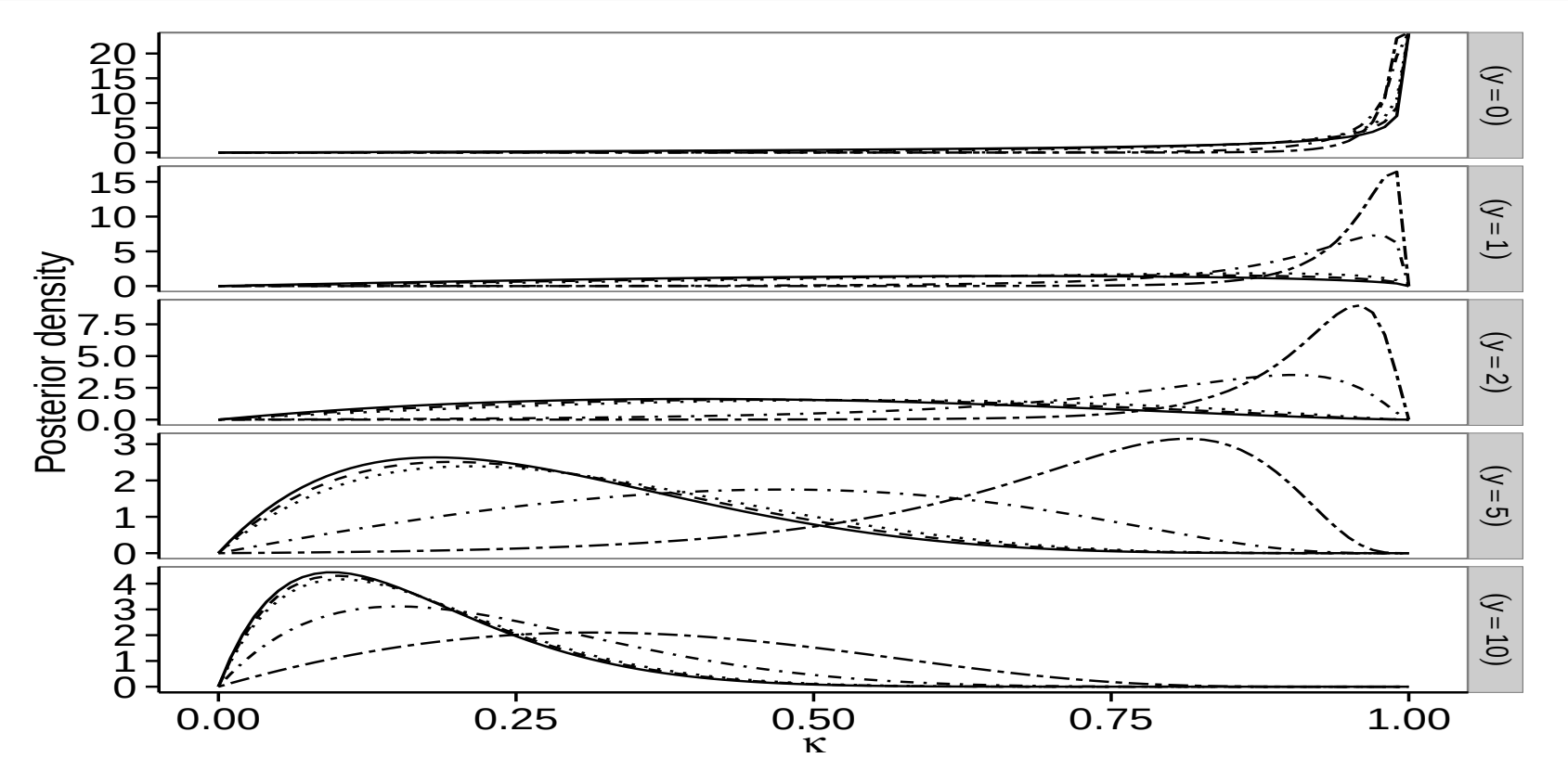
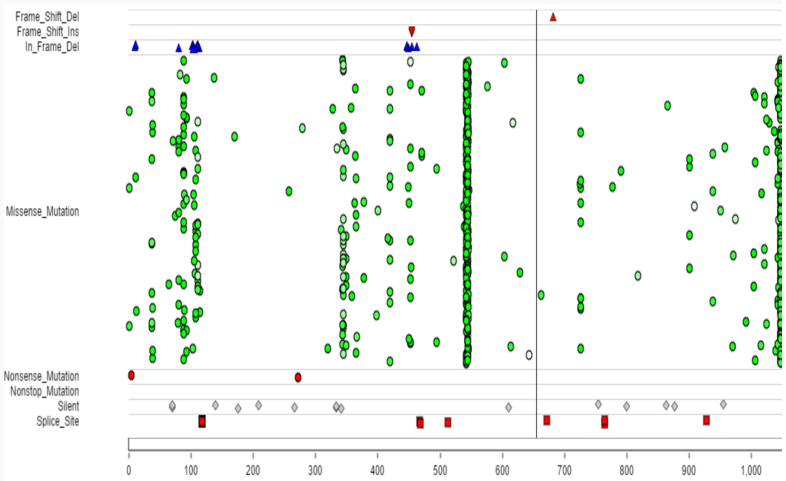


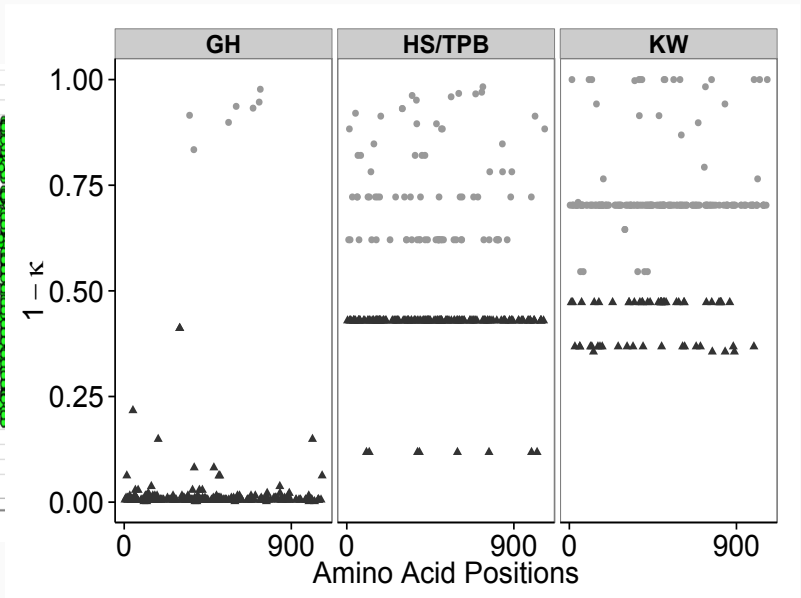
Figure 1: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), $\gamma = 10$ (two-dash).

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.



(a) Distribution of mutations by type of tumor on the gene 'PIK3CA'



(b) Comparison of Different Shrinkage Profiles

▸ Bayesian Duality ▸ Skip to Shrinkage on Simplex

Shrinkage on simplex [2]

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 $\boldsymbol{\pi}$.
2. Model a general dependence structure.

[2] ongoing work with 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 $\text{Dir}(\alpha)$ can be inflexible for modeling sparse probabilities.
- $\text{Dir}(\alpha)$ density can be tuned to concentrate near 1-sparse vectors, but it is difficult to favor π 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.
- $f = \text{Be}\left(\frac{\alpha}{K},\ \alpha(1 - \frac{k}{K})\right) \Rightarrow \boldsymbol{\pi} \sim \text{Dir}(\alpha)$.

Constructive Definition

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

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

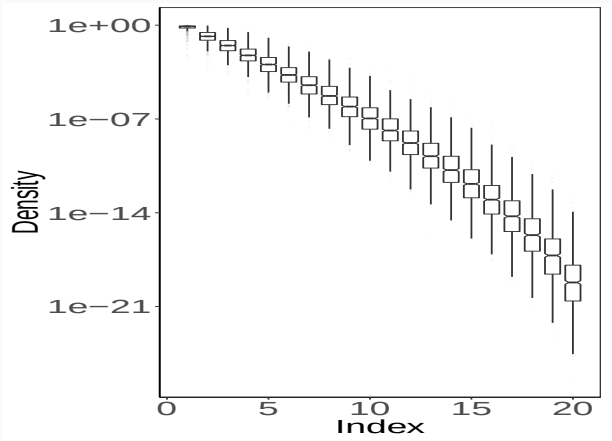
$$\pi_1 = Z_1,\ \pi_k = Z_k \prod_{l=1}^{k-1}(1 - Z_l),\ k = 1, \ldots, K-1,$$

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

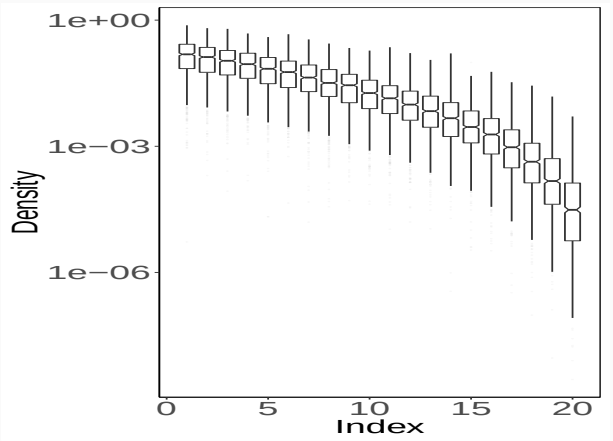
- $f(\cdot)$ is a density supported on the unit interval.
- $f = \text{Be}\left(\frac{\alpha}{K},\ \alpha(1 - \frac{k}{K})\right) \Rightarrow \boldsymbol{\pi} \sim \text{Dir}(\alpha)$.
- Idea: Use global-local shrinkage prior for $f(\cdot)$, e.g. Horseshoe or Gauss-hypergeometric prior.

Sparse Generalized Dirichlet

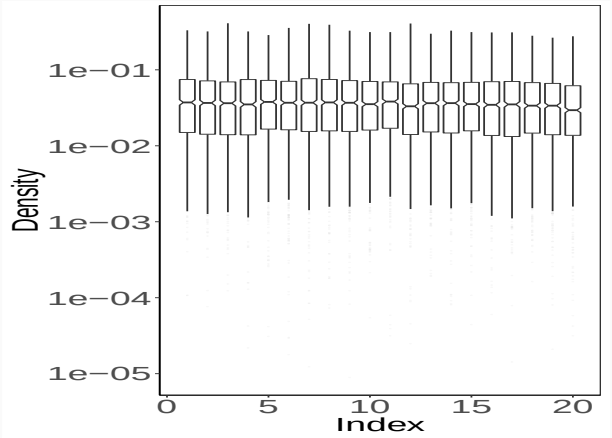
Sparsity prior: $\mathrm{GH}(\alpha/K,\ \alpha(1-k/K), \phi)$: induces an ordering and converges to a standard $\mathrm{Dir}(\alpha)$ distribution as $\phi \to 1$.



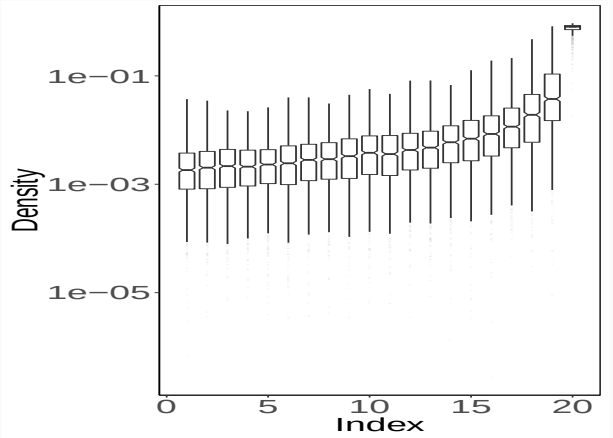
(a) ($\phi = 0.01$)



(b) ($\phi = 0.2$)



(c) ($\phi = 0.95$)



(d) ($\phi = 20.0$)

Density

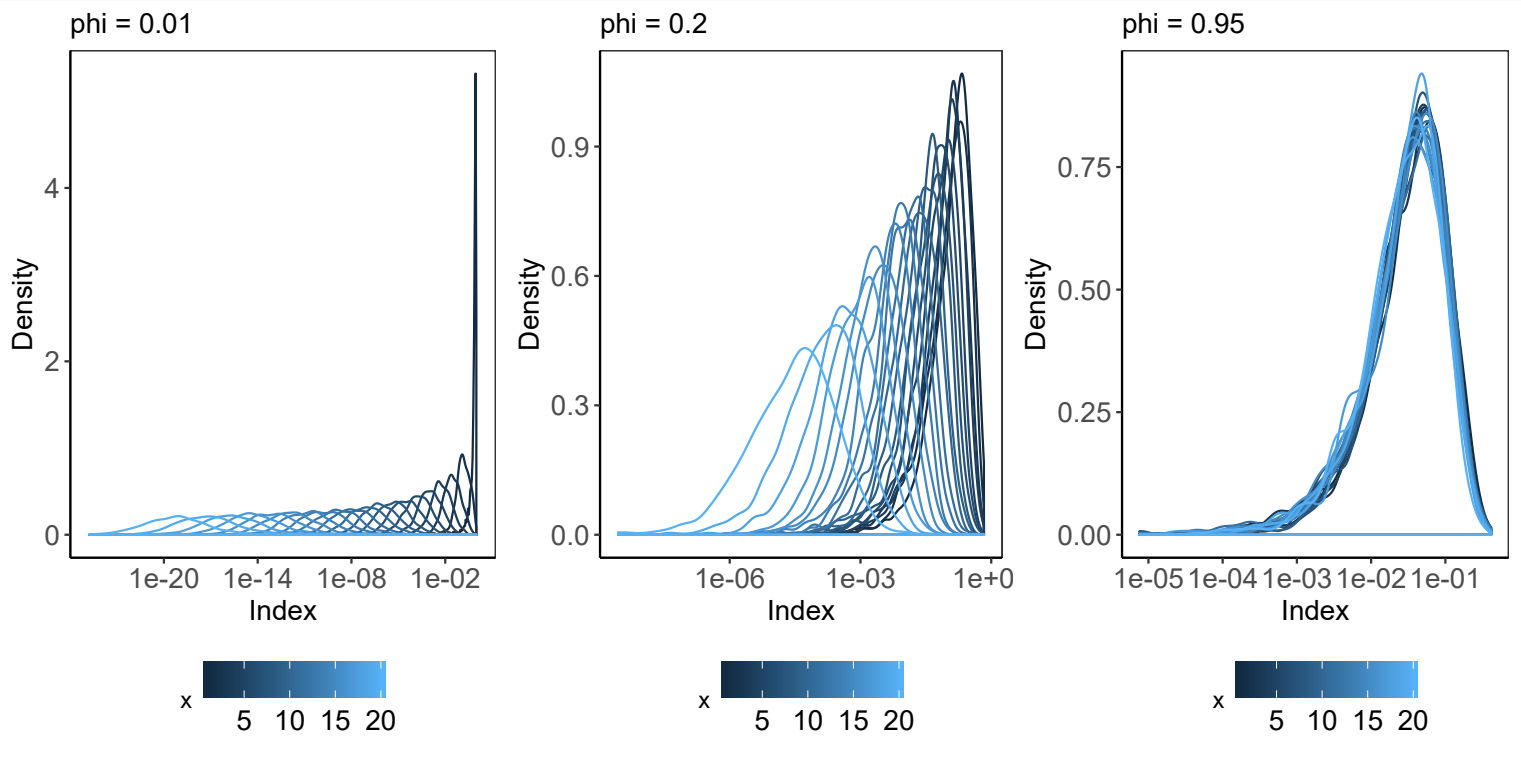


Figure 4: Density estimates for 1,000 draws of π_i from the type 1 $\mathcal{SGD}$ distribution for $K = 20$ when $\phi \in \{0.01, 0.2, 0.95\}$ (left to right). The varying colors indicate densities different components $i \in \{1, \ldots, K\}$.

Form of Density

Use GH priors in stick-breaking: Sparse Generalized Dirichlet distribution with parameters $\mathbf{a}, \mathbf{b}, \phi$ and γ,

$$\boldsymbol{\pi} \sim \mathcal{SGD}(\mathbf{a}, \mathbf{b}, \boldsymbol{\phi}, \gamma) \Leftrightarrow Z_i \sim f = \mathrm{GH}\left(a_i, b_i, \phi, \gamma\right), i = 1, \ldots, K-1. \tag{1}$$

Proposition

Let $\boldsymbol{\pi} \sim \mathcal{SGD}(\mathbf{a}, \mathbf{b}, \phi, \gamma)$ with $a_i = \alpha(1 - i/K)$ and $b_i = \alpha/K$, $\gamma_i = a_i + b_i$, and $\phi_i - 1 = (\phi - 1)(1 - S_{i-1})$ for $\alpha, \phi > 0$, then the density for $\boldsymbol{\pi}$ is:

$$f(\boldsymbol{\pi} \mid \alpha, \boldsymbol{\phi}) \propto \pi_1^{\frac{\alpha}{K}-1} \cdots \pi_K^{\frac{\alpha}{K}-1} \prod_{i=1}^{K-1} \left\{1 + (\phi - 1)\pi_i\right\}^{-\{\alpha(1-\frac{i}{K})-1\}}, \tag{2}$$

Clearly, $\phi = 1$ will reduce this to a Dirichlet.

Sparse Generalized Dirichlet prior is conjugate to the multinomial likelihood.

Simulation Study: Sparse π_0

Sparse $\pi_0 = (1/2, \ldots, 1/2^{K-1}, 1/2^{K-1})$.

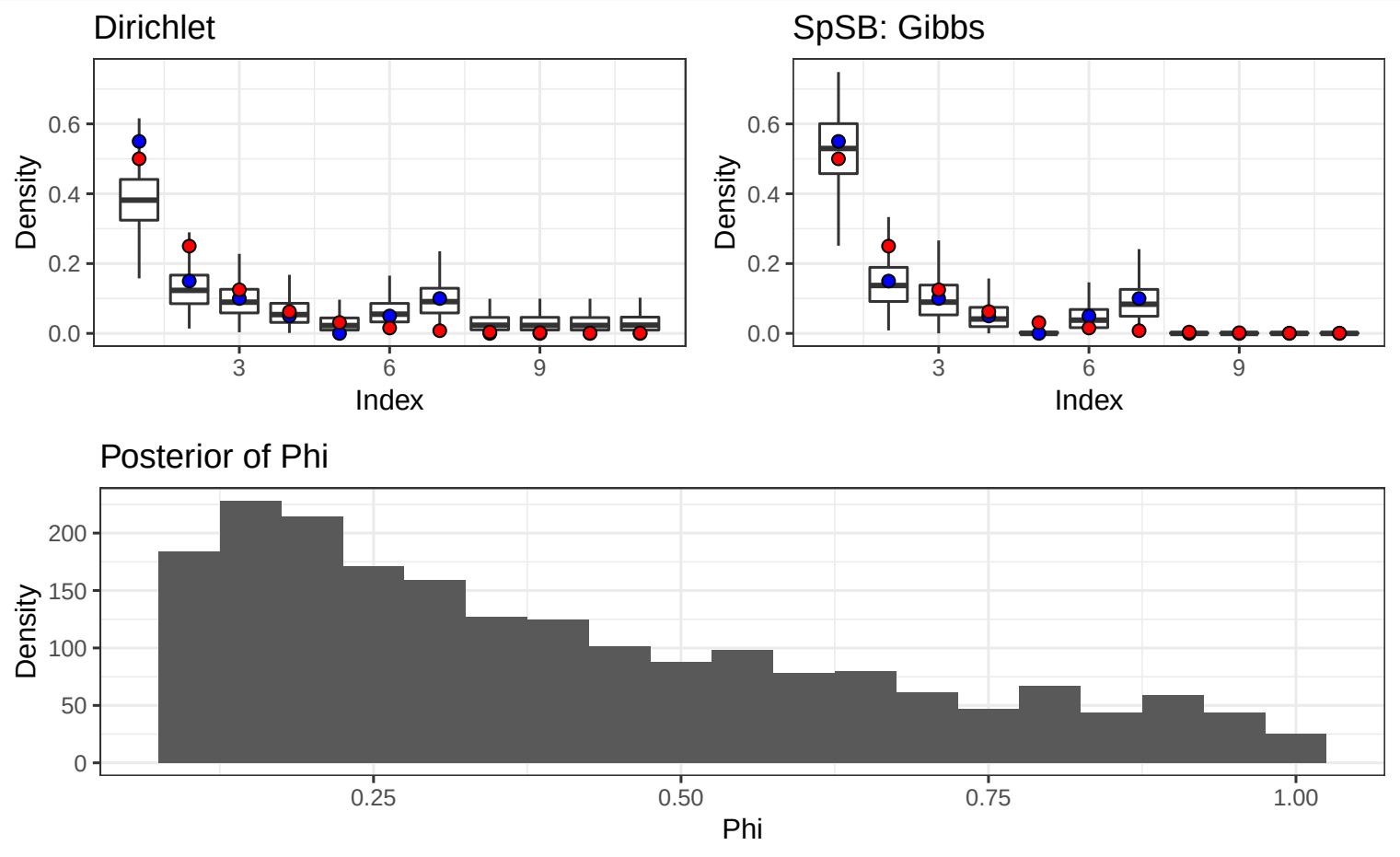
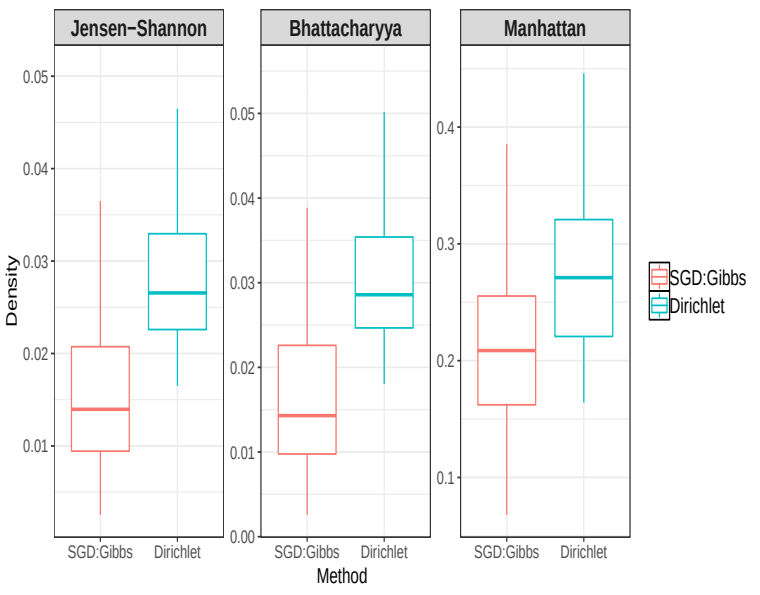
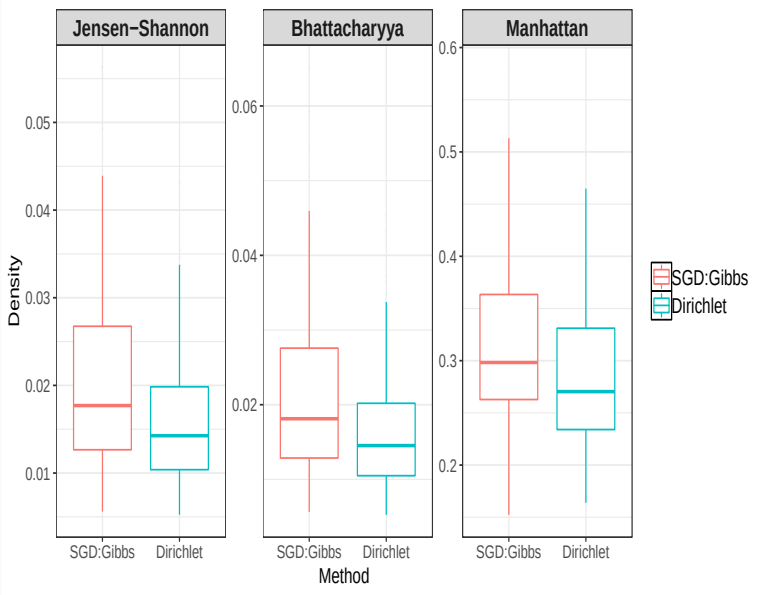


Figure 5: Top row: Posterior distribution of π given y for sparse π_0 (**??**) under the candidate priors. The blue and red dots denote the observed proportions and the true π_0 values. Bottom row: Histogram of posterior samples of ϕ

Repeated Simulations



(a) Different divergence measures for sparse π_0.



(b) Different divergence measures for uniform π_0.

The Sparse Generalized Dirichlet outperforms Dirichlet when π_0 is 'sparse' and has no advantage over Dirichlet when π_0 is uniform.

Theory: Neutrality & Tail-contraction

- 'Neutrality' : natural way of characterizing independence for compositional data satisfying the unity-sum constraint.
- Let $Z_j = \pi_j / (\sum_{i \geq j} \pi_i)$: representing the proportion of the remaining interval to be assigned to the next variable π_j.
- Complete neutrality $\equiv Z_1, Z_2, \ldots, Z_K$ are mutually independent [Connor and Mosimann, 1969].
- The $\mathcal{SGD}$ distribution satisfies complete neutrality.
- We can show that if $\phi \to 0$, the tail of $\boldsymbol{\pi} \sim \mathcal{SGD}$ distribution would place most of its density in a vanishing neighborhood around zero.

Modeling Community Types

Sparsity-favouring distribution on species abundance.

$$\mathbf{Y}_i \mid \pi_i \sim \text{Multinomial}(y_{i+} \mid \pi_i),\ i = 1, \ldots, M,$$
$$\pi_i \mid c_i = h \sim \mathcal{SGD}_1(\alpha_i \mathbf{1}, \phi_h), h = 1, \ldots, K,$$
$$\log(\alpha_i \mid c_j = h) = \beta_h^T \mathbf{X}_i + \epsilon_i,\ h = 1, \ldots, K,$$
$$\beta_h \sim \mathcal{N}(\xi, \Psi), \text{ and } \phi_h \sim \text{U}(\delta, 1),\ h = 1, \ldots, K,$$
$$C_i \sim \text{Categorical}(\boldsymbol{\lambda}), \text{ where } \lambda_k = P(c_i = k),$$
$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K) \sim \text{Dir}(\gamma)\ .$$

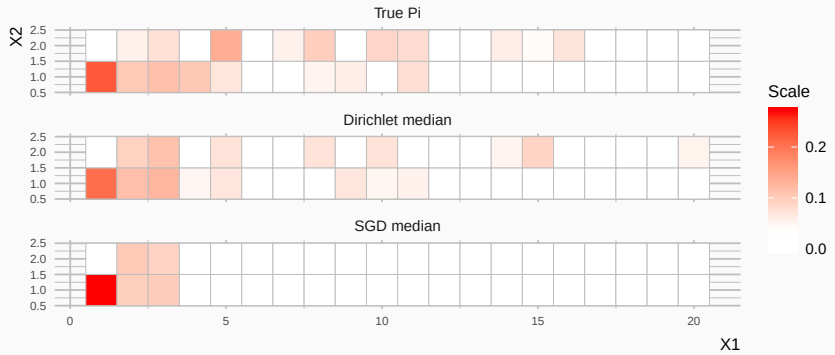
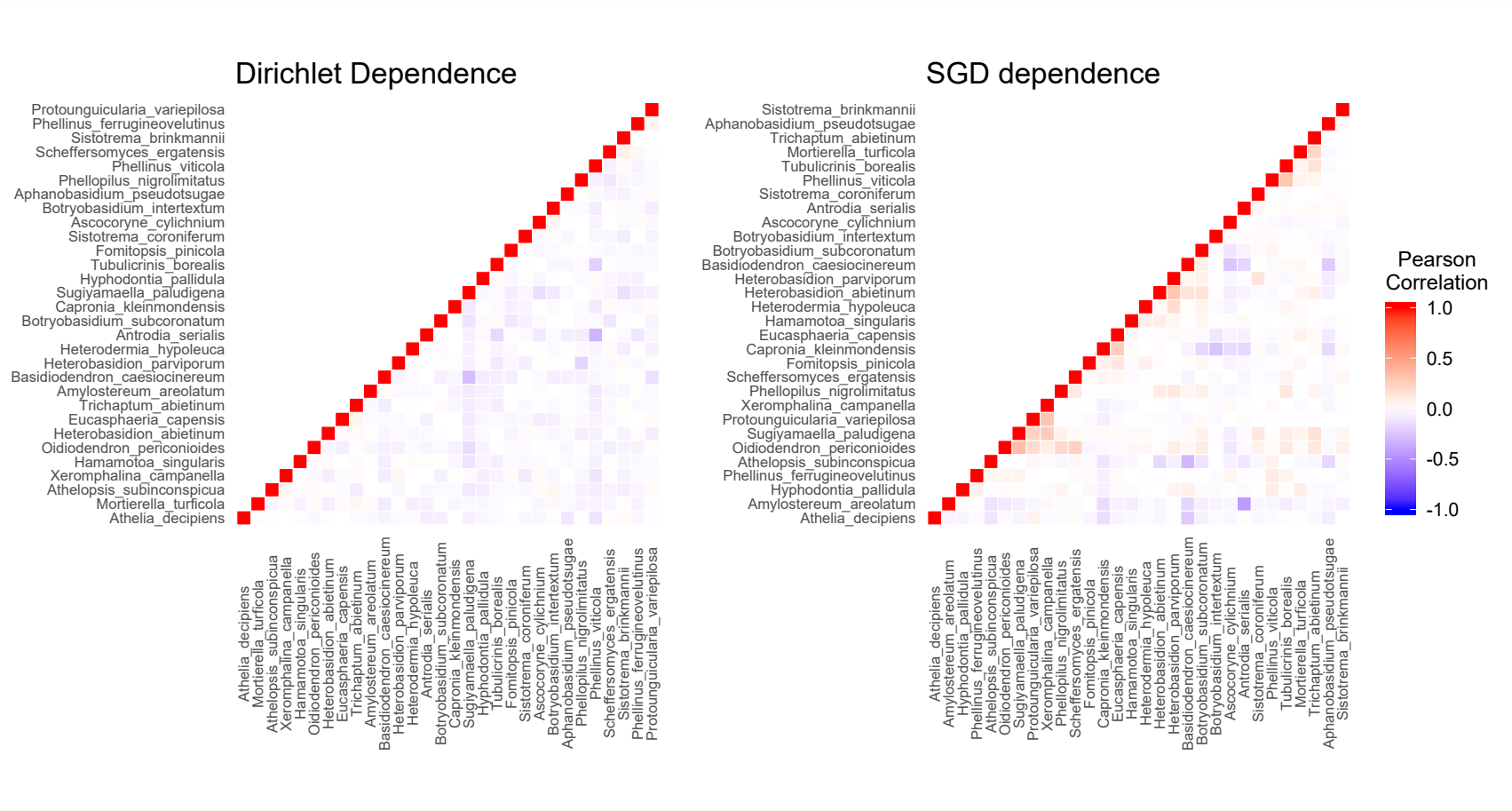


Figure 7: Posterior mean estimate for Π under Dirichlet and Sparse Generalized Dirichlet prior

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

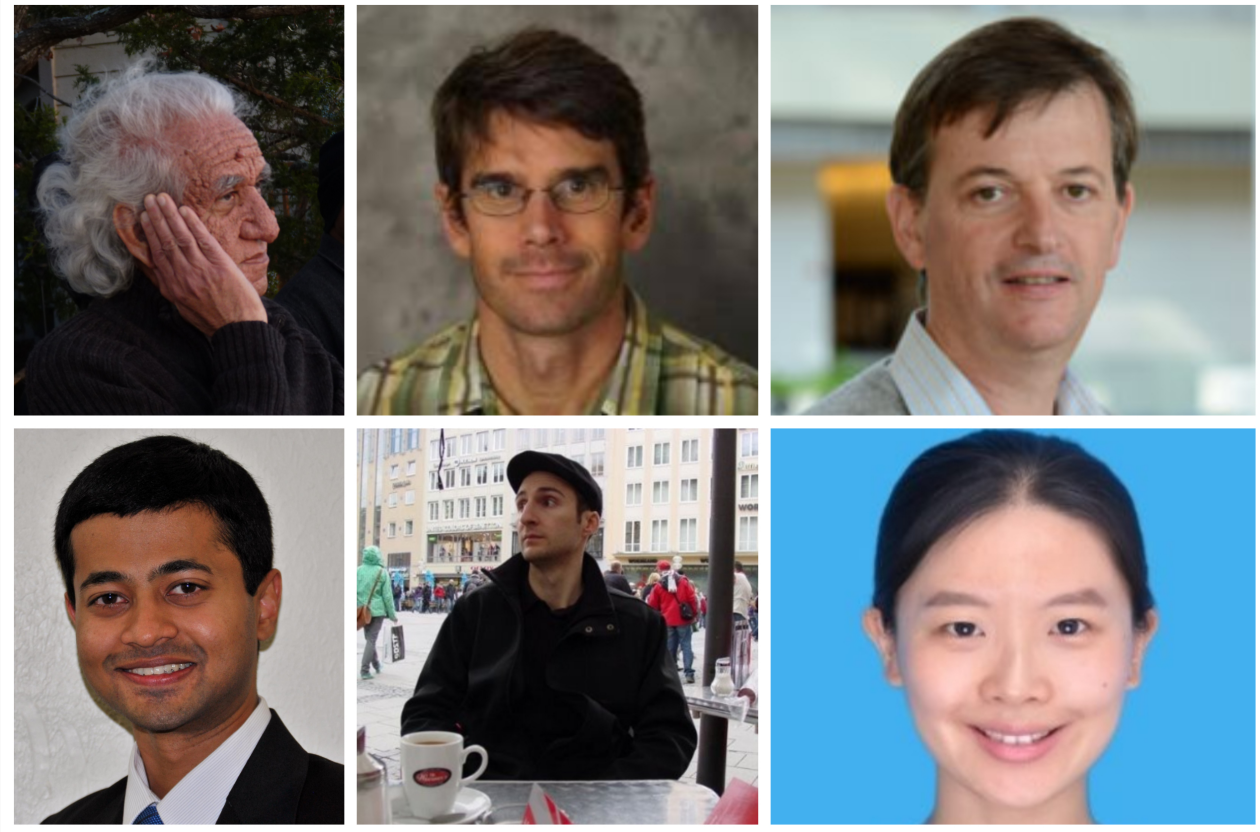
Summary & Directions

- Use G-L priors for stick-breaking construction of Dirichlet: Shrinkage on Simplex !
  Next steps:
- Hierarchical Extension?
- Usage in network models and topic modeling
- Sparsity of species. Affinity / dependence?
- Testing framework.

References

- 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. (2019). Horseshoe regularization for feature subset selection. *Accepted*, *Sankhya B*. [preprint]
- 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!



Supplementary Slides for HSlike

Bayesian Regularization: A Useful Duality

- Regularization leads to an optimization problem of the form

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ l(\mathbf{y} \mid \boldsymbol{\theta}) + \text{pen}_\lambda(\boldsymbol{\theta}) \right\} \ . \tag{3}$$

- Probabilistic approach leads to a Bayesian hierarchical model

$$p(y \mid \boldsymbol{\theta}) \propto \exp\{-l(y \mid \boldsymbol{\theta})\} \ , \quad p_\lambda(\boldsymbol{\theta}) \propto \exp\{-\text{pen}_\lambda(\boldsymbol{\theta})\}. \tag{4}$$

- Regularized estimate $\equiv$ Posterior mode.
- e.g. ℓ_2 penalty (Ridge) $=$ [Hoerl and Kennard, 1970] Gaussian prior, and
- ℓ_1 penalty (Lasso) $=$ double-exponential / Laplace prior [Tibshirani, 1996].

Lack of heavy tails : Bayesian Lasso

- The posterior mode is identical to Lasso solution, should inherit optimal features of Lasso !
- Castillo et al. [2015]: the full Lasso posterior distribution does not contract **at the same speed as the posterior mode** $\Rightarrow$ Poor uncertainty quantification.
- Insufficient shrinkage by Laplace prior: Polson and Scott [2010b], Datta and Ghosh [2013] - motivates Horseshoe.

Regularized Regression

- Convex penalties: Pros: unique solution, efficient computation and straightforward theory.
- Cons: Bias $\Rightarrow$ poor estimation error, tends to select a denser model (Mazumder et al. 2012), strong assumptions on the design matrix (coherence/irrepresentability).
- Ideal: ℓ_0: NP-hard.
- Non-convex penalty (e.g ℓ_γ for $\gamma \in (0, 1)$), sparser model, need weaker coherence condition, low SNR.
- Harder to fit: convex optimization tools do not apply, nor is a globally optimal solution guaranteed.

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 (28) recovers the bounds in (**??**).

Horseshoe-like prior

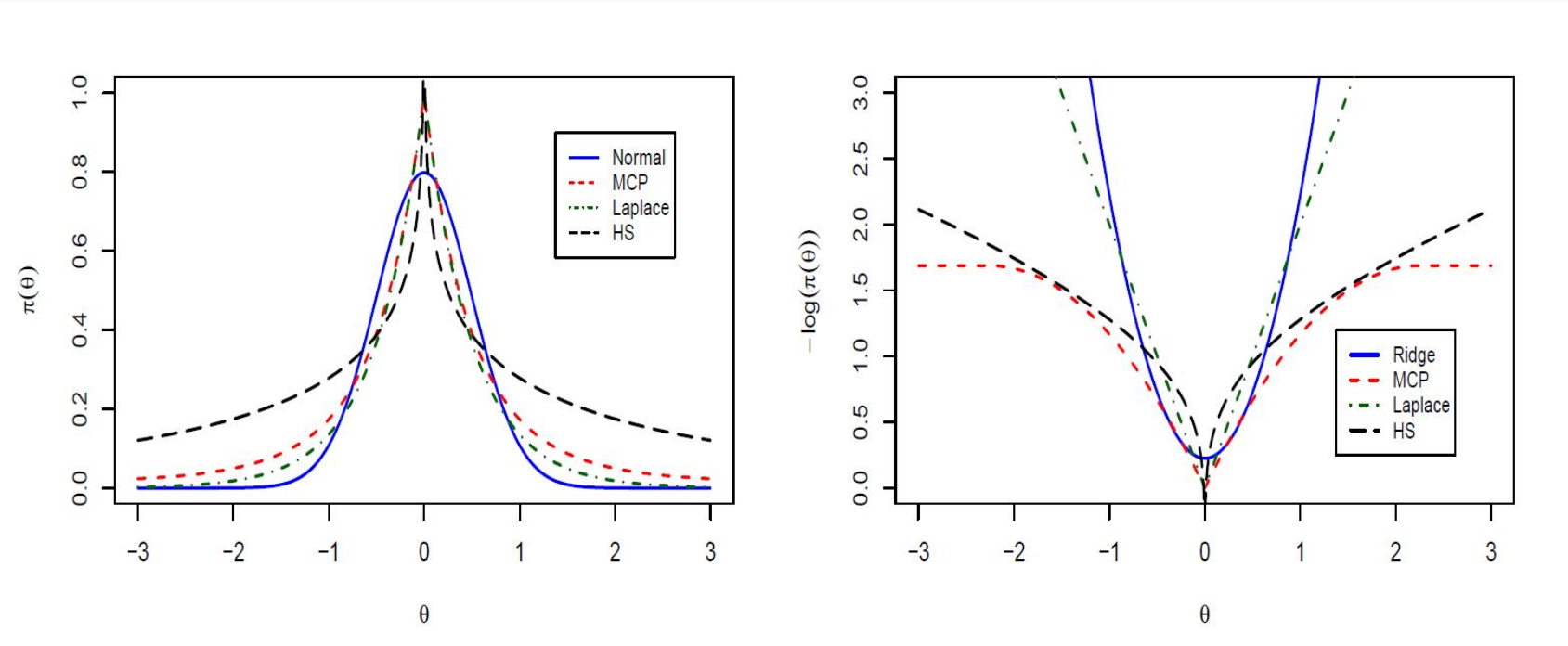


Figure 9: (a) Prior density (b) Induced Penalty

- 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 [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:

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

- Normal scale mixture:

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

EM algorithm

- Regression:

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

- E-step: compute posterior expectations $\tilde{u}_i = E(u_i \mid \theta_i, y_i, a)$,
- M-step: maximize the full posterior jointly [3] in (θ, a) with $u_i \mapsto \tilde{u}_i$.
- EM: $(t+1)^{\text{th}}$ iteration for $t \geq 0$ is:

$$\hat{a}^{(t+1)} \mid \hat{\theta}_1^{(t)}, \ldots, \hat{\theta}_p^{(t)} = \frac{\{\hat{a}^{(t)}\}^{3/2}}{p\pi} \sum_{i=1}^{p} \left(\frac{1}{\{\hat{\theta}_i^{(t)}\}^2 + \hat{a}^{(t)}} \right),$$

$$\hat{\theta}^{(t+1)} \mid \hat{a}^{(t+1)} = \underbrace{\left\{ X^T X + \operatorname{diag}\left(\frac{2\tilde{u}_i^{(t)}}{\hat{a}^{(t+1)}} \right) \right\}^{-1}}_{\text{inverse of a } p \times p \text{ matrix}} X^T y,$$

[3] Easy to maximize $\theta \mid a$ and $a \mid \theta$ because of Normal scale mixture representaion

MCMC

- Hierarchy for horseshoe-like prior:

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

- We need full conditionals for Gibbs sampling !
- Block-update all local parameters to increase efficiency.

MCMC

- Hierarchy for horseshoe-like prior:

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

- We need full conditionals for Gibbs sampling !
- Block-update all local parameters to increase efficiency.
- **We need one more trick!**

Scale mixture

- Reparametrize ($t_i^2 = 2u_i$ and $\tau^2 = a$):

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

- This $p(t_i)$ is the standard Slash-Normal density that can be written as a Normal variance mixture with a $\text{Pareto}(\frac{1}{2})$ mixing density.
- Full scale mixture representation for horseshoe-like:

$$(\theta_i \mid t_i, \tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{t_i^2}\right), (t_i \mid s_i) \sim \mathcal{N}(0, s_i), \; s_i \sim \text{Pareto}(1/2).$$

- Gibbs sampling follows from this !

Simulation: Regression

- $\mathbf{Y} = \mathbf{X}_{n\times p}\boldsymbol{\theta} + \boldsymbol{\epsilon},\ n = 70, p = 350.$
- True $\boldsymbol{\theta} = (\underbrace{3,\ldots,3}_{q_n=10}, \underbrace{-3,\ldots,-3}_{q_n=10}, \overbrace{0,\ldots,0}^{n-2q_n=330})$.
- The matrix of predictors $X \in \mathbb{R}^{n\times p}$ drawn from i.i.d. standard normals.
- HS posterior mode has the second best performance in detection of zeros.
- SCAD detects the highest number of true zeros correctly, but poor performance in detection of non-zeros (7 out of 20) and a poor SSE.
- HS posterior mode gives sparser solution compared to MCP and Lasso.

	HS-mode	HS-mean	SCAD	MCP	Lasso
SSE	91.03	44.35	143.2	42.55	66.93
TN	302	NA	323	276	292
TP	18	NA	7	20	20
Time	0.248	14.978	0.226	0.561	0.177

Table: Sum of squared error (SSE), true nulls and non-nulls (TN & TP) and time in s. (TIME).

Summary - Horseshoe-like

- Horseshoe-like: Global-local prior + non-convex penalty.
- Scale mixture allows for MCMC for full Bayes and EM and LLA algorithms for MAP point estimates.
- Next steps:
  1. Can we investigate the penalty produced by a general global-local prior?
  2. Can we extend the HS penalty for GLM?
  3. Posterior mean optimality transfer to posterior mode optimality?

Theorem (Gribonval, 2011)
For any prior $p(\theta)$, the posterior mean can be interpreted as a MAP with some prior $C\exp(-\phi(\theta))$. Vice-versa, for certain penalties $\phi(\theta)$, the penalized least squares estimate is the posterior mean with a certain prior $p(\theta)$. In general we have $p(\theta) \neq C\exp(-\phi(\theta))$.

More details: Horseshoe Regularization for Feature Subset Selection: Bhadra, Datta, Polson & Willard. arXiv: 1702.07400.

▸ Shrinkage on Simplex ▸ References

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 1: Implementations of Horseshoe and Other Shrinkage Priors

Implementation (Package/URL)	Authors
R package: `monomvn`	Gramacy et al. [2010]
R code in paper	Scott [2010]
R package: `horseshoe`	van der Pas et al. [2016b]
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]

Applications/Extensions of Lasso

Table 2: A few regularization methods

Method	Authors
Adaptive Lasso	Zou [2006]
Compressive sensing	Donoho [2006], Candes [2008]
Dantzig selector	Candes and Tao [2007]
Elastic net	Zou and Hastie [2005]
Fused Lasso	Tibshirani et al. [2005]
Generalized Lasso	Tibshirani and Taylor [2011]
Graphical Lasso	Friedman et al. [2008]
Grouped Lasso	Yuan and Lin [2006]
Hierarchical interaction models	Bien et al. [2013]
Matrix completion	Candès and Tao [2010], Mazumder et al. [2010]
Multivariate methods	Jolliffe et al. [2003], Witten et al. [2009]
Near-isotonic regression	Tibshirani et al. [2011]
Square Root Lasso	Belloni et al. [2011]
Scaled Lasso	Sun and Zhang [2012]
Minimum concave penalty	Zhang [2010]
SparseNet	Mazumder et al. [2012]

Applications/Extensions of Horseshoe

Table 3: Applications of the horseshoe prior

Application	Authors
Fadeout method for mean-field variational inference under non-centered parameterizations and stochastic variational inference for undirected graphical model.	Ingraham and Marks [2016]
Linear regression for Causal inference and Instrumental variable models	Hahn and Lopes [2014], Hahn et al. [2016]
Multiclass prediction using DOLDA (Diagonally orthant Latent Dirichlet Allocation)	Magnusson et al. [2016]
Mendelian Randomization to detect causal effects of interest	Berzuini et al. [2016]
Locally adaptive nonparametric curve fitting with shrinkage prior Markov random field (SPMRF)	Faulkner and Minin [2015]
Quasi-Sparse Count Data	Datta and Dunson [2016]
Variable Selection under the projection predictive framework	Piironen and Vehtari [2015]

Supplementary Slides for Simplex Shrinkage

Computation: Conjugacy

Lemma

Suppose the prior distribution on $\boldsymbol{\pi}$ *is a* $\mathcal{SGD}(\mathbf{a}, \mathbf{b}, \boldsymbol{\phi}, \gamma)$*, and the sampling model* $\mathbf{Y} \mid \boldsymbol{\pi}$ *follows a multinomial distribution with* y_j*,* $j = 1, 2, \ldots, K$ *as defined above. Then the joint posterior of* $\boldsymbol{\pi} \mid \mathbf{Y}$ *is also a Sparse Generalized Dirichlet distribution with updated shape parameters* $\mathbf{a}', \mathbf{b}'$ *and the same parameters* $\mathbf{fl}$ *and* $\boldsymbol{\phi}$*.*

$$
\begin{aligned}
&[\boldsymbol{\pi} \mid \mathbf{Y}] \sim \mathcal{SGD}(\mathbf{a}', \mathbf{b}', \gamma, \boldsymbol{\phi}), \\
&\text{where } a_i' = a_i + y_i, \;\; b_i' = b_i + \sum_{j>i} y_j, \;\; i = 1, \ldots, K-1.
\end{aligned} \tag{5}
$$

Computation: Gibbs-Slice

Easy to write a Gibbs sampler using a data augmentation approach.

$$\boldsymbol{\pi} \mid \text{rest} \sim \mathcal{SGD}(\mathbf{a}', \mathbf{b}', \mathbf{fl}, \boldsymbol{\phi}), \tag{6}$$

$$\text{where } a_i' = \alpha(1 - i/K) + y_i,\ b_i' = \alpha/K + \sum_{j>i} y_j, \gamma = \alpha(1 - \frac{i-1}{K})$$

$$\phi \mid \text{rest} \sim \mathcal{E}xp\left(-\sum_{i=1}^{K-1} \frac{\pi_i}{1 - S_i}\right) \tag{7}$$

$$\omega_i \mid \text{rest} \sim \mathcal{G}\left(\left(1 + (\phi - 1)\frac{\pi_i}{1 - S_i}\right), a_{i-1} = \alpha(1 - \frac{i-1}{K})\right) \tag{8}$$

Repeated Simulations: Sparse

Method	Mean	Median	SD
JSD-$\mathcal{SGD}$	0.0158	0.0140	0.0084
JSD-Dir	0.0283	0.0265	0.0073
Bhatta-$\mathcal{SGD}$	0.0166	0.0143	0.0092
Bhatta-Dir	0.0307	0.0286	0.0080
Manhattan-$\mathcal{SGD}$	0.2150	0.2086	0.0774
Manhattan-Dir	0.2779	0.2711	0.0683

Comparison between the estimation accuracies by $\mathcal{SGD}$ prior and the Dirichlet prior for data generated from a Multinomial$(n = 50, \pi_0)$ with π_0 being the sparse geometric sequence in (??) over 100 simulations. JSD: Jensen-Shannon divergence.

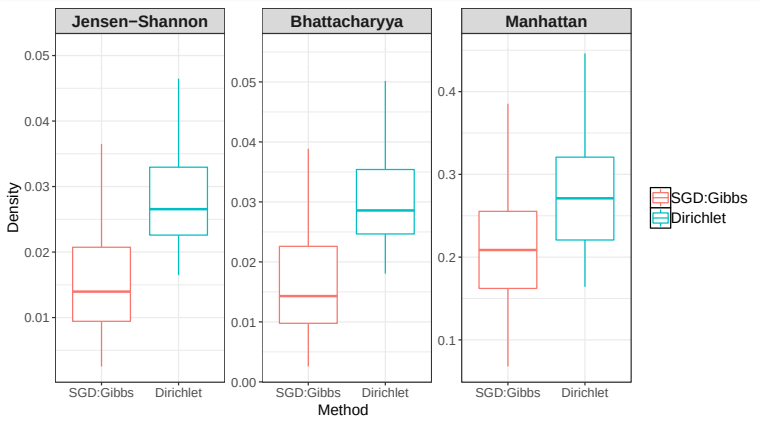


Figure 10: Comparison of the different divergence measure distributions for the sparse π_0.

Uniform π_0

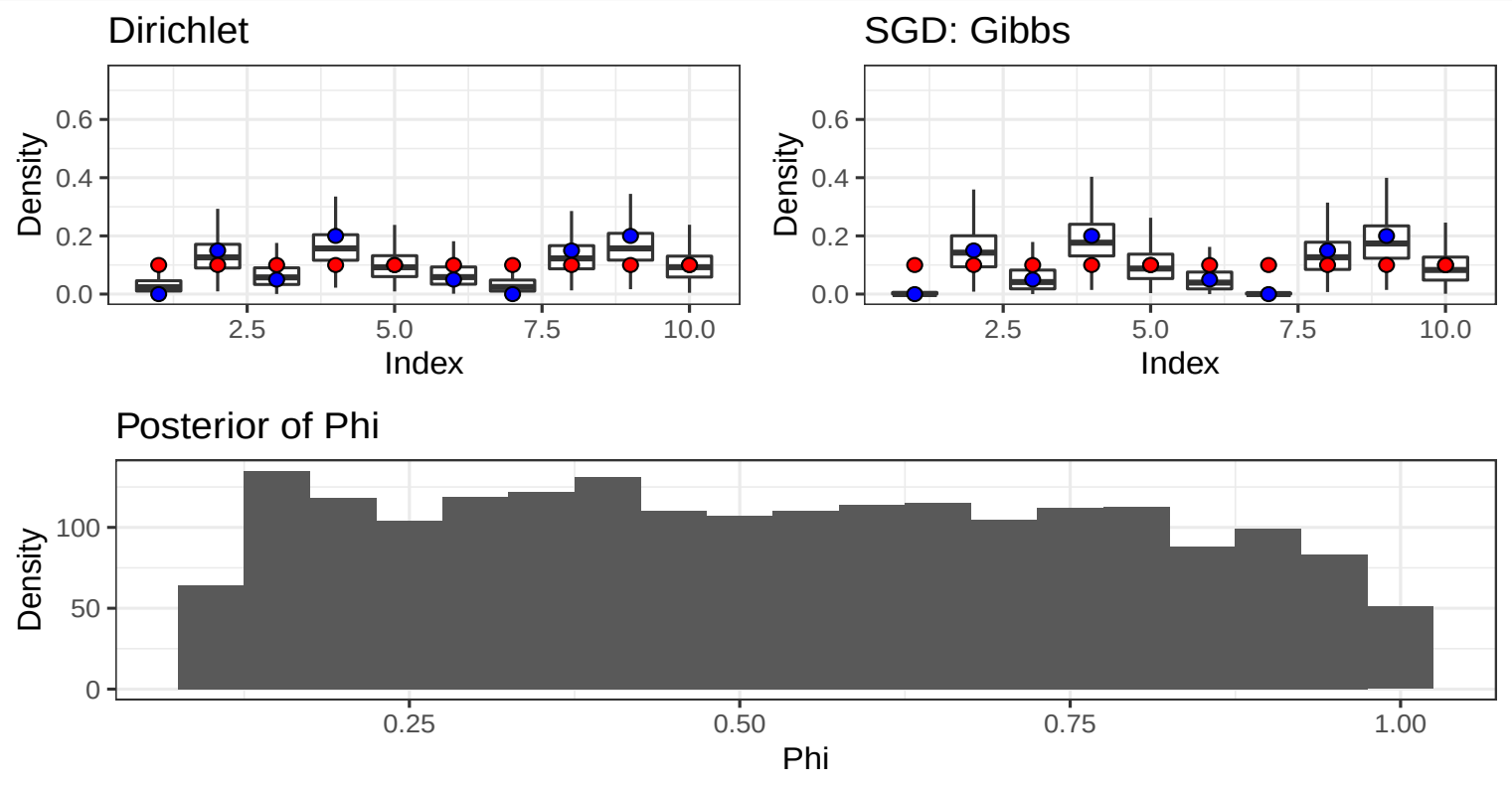


Figure 11: Top row: Posterior distribution of π given y for unifrom π_0 (**??**) under the candidate priors: the symmetric Dir(α), and $\mathcal{SGD}$ prior with ϕ learned via a full Bayes approach. The blue and red dots denote the observed and the true π_0 values. Bottom row: Histogram of posterior samples of ϕ.

Repeated Simulations: Uniform

Method	Mean	Median	SD
JSD-$\mathcal{SGD}$	0.0208	0.0177	0.0112
JSD-Dir	0.0160	0.0143	0.0080
Bhatta-$\mathcal{SGD}$	0.0216	0.0181	0.0123
Bhatta-Dir	0.0163	0.0145	0.0083
Manhattan-$\mathcal{SGD}$	0.3169	0.2983	0.0860
Manhattan-Dir	0.2795	0.2704	0.0752

Comparison between the estimation accuracies by Sparse Generalized Dirichlet prior and the Dirichlet prior for observations generated from (??) over 100 simulations. JSD: Jensen-Shannon divergence.

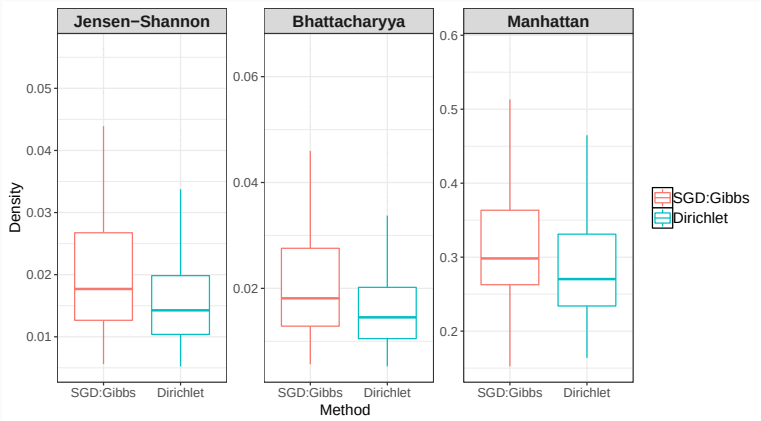


Figure 12: Comparison of the different divergence measure distributions for the sparse π_0.

Can we learn ϕ?

The true composition vector was drawn from the Sparse Generalized Dirichlet distribution proposed here (12), with a fixed sparsity parameter $\phi = 0.1$.

$$
\begin{aligned}
(y_1, \ldots, y_K) &\sim \text{Multinomial}(n, \pi_1, \ldots, \pi_K), \\
\text{Sparse Generalized Dirichlet } \boldsymbol{\pi} &\sim \mathcal{SSGD}(\alpha = 1, \phi = 0.1), \qquad (9)
\end{aligned}
$$

Learning ϕ

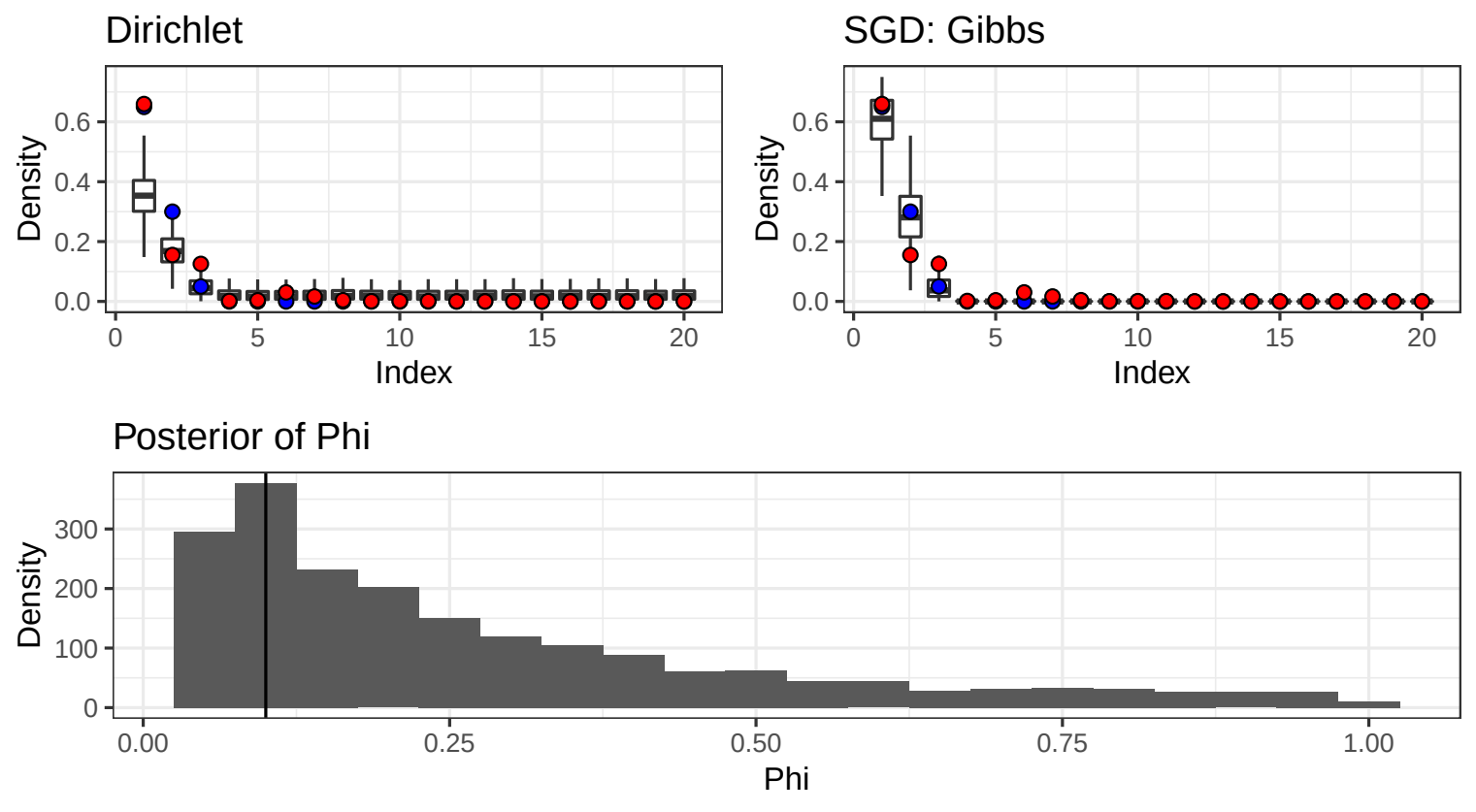


Figure 13: Top row: Posterior distribution of π given y for $\pi_0 \sim \mathcal{SSGD}(\alpha = 1, \phi = 0.1)$ (9), under the candidate priors: the symmetric $\text{Dir}(\alpha)$, and $\mathcal{SGD}$ prior with ϕ learned via a full Bayes approach. The blue and red dots denote the observed and the true π_0 values. Bottom row: Histogram of posterior samples of ϕ along with the true value $\phi = 0.1$.

Theory: Neutrality

- 'Neutrality' : natural way of characterizing independence for compositional data satisfying the unity-sum constraint.
- Consider the composition vector $\boldsymbol{\pi}$ and $Z_j = \pi_j/(1 - S_{j-1})$: representing the proportion of the remaining interval to be assigned to the next variable π_j.
- Connor and Mosimann [1969] showed that complete neutrality is equivalent to the property $Z_1, Z_2, \ldots, Z_K$ are mutually independent.
- Natural requirement in many contexts such as Bayesian life-table analysis where conditional failure probabilities are assumed to be independent, or in analyzing chemical compositions.

Theory: Neutrality

Lemma

Suppose, $(\pi_1, \ldots, \pi_K) \sim \mathcal{SGD}(\mathbf{a}^K, \mathbf{b}^K, \phi^K, \gamma^K)$*, where the super-script denotes the size of the hyper-parameters, i.e.* $\mathbf{a}^k = (a_1, \ldots, a_k)$ *for* $1 \leq k \leq K$ *and so on, then it turns out that:*

$$Z_j \doteq \frac{\pi_j}{1 - \sum_{i<j} \pi_i} \Big| \{\pi_1, \ldots, \pi_{j-1}\} \sim GH(a_j, b_j, \phi_j, \gamma_j)$$

$$(\pi_1, \ldots, \pi_{j-1}) \sim \mathcal{SGD}(\mathbf{a}^{j-1}, \mathbf{b}^{j-1}, \boldsymbol{\phi}^{j-1}, \gamma^{j-1}) \quad \text{for} \quad 2 \leq j \leq K.$$