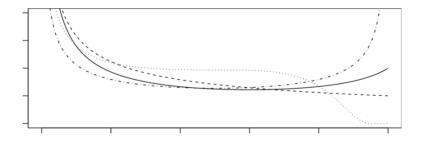
Scalable MCMC for Bayes Shrinkage Priors

Paulo Orenstein July 2, 2018

Stanford University



Joint work with James Johndrow and Anirban Bhattacharya

Introduction	Model	Computation	Results	Conclusion
Introduction				
Consider 1	:he high-dimensional s	setting: predict a vector	$y \in \mathbb{R}^n$ from a set	of

▶ Consider the high-dimensional setting: predict a vector $y \in \mathbb{R}^n$ from a set of features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.

Introduction	Model	Computation		Results	Conclusion
Introduction					
Consider	the high-dimensior	nal setting: pr	edict a vector	$y \in \mathbb{R}^n$ from a set	et of

features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.

Assume a sparse Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with $\beta_j = 0$ for many *j*.

Introductio					
Introd	uction				
	Consider the	high-dimensional se	tting: predict a vector	$y \in \mathbb{R}^n$ from a set	of

- Consider the high-dimensional setting: predict a vector $y \in \mathbb{R}^n$ from a set of features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.
- Assume a sparse Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with $\beta_j = 0$ for many *j*.

▶ How can we perform prediction and inference?

Introduction				
Introduc	tion			
•	Consider the high-dimensiona	l setting: predict a	vector $y \in \mathbb{R}^n$ from	a set of

- Consider the high-dimensional setting: predict a vector $y \in \mathbb{R}^n$ from a set o features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.
- Assume a sparse Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with $\beta_j = 0$ for many *j*.

How can we perform prediction and inference?

Lasso

Introduction				
Introduc	tion			
	Consider the high-dimensiona	l setting: predict a	vector $y \in \mathbb{R}^n$ from	a set of

- ▶ Consider the high-dimensional setting: predict a vector $y \in \mathbb{R}^n$ from a set of features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.
- Assume a sparse Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with $\beta_j = 0$ for many *j*.

- How can we perform prediction and inference?
 - Lasso
 - Point mass mixture prior

Introduction				
Introduction				
Consi	der the high-dimensio	onal setting: predict a	vector $y \in \mathbb{R}^n$ from	a set of

- features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.
- Assume a sparse Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with $\beta_i = 0$ for many *j*.

- How can we perform prediction and inference?
 - Lasso, but: convex relaxation; one parameter for sparsity and shrinkage
 - Point mass mixture prior

Introduction					
Introduc	tion				
millouuc	tion				
	Consider the high	gh-dimensional setting:	r predict a vector y	$\mathcal{V} \in \mathbb{R}^n$ from a set of	of

- features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$.
- Assume a sparse Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with $\beta_i = 0$ for many *i*.

- How can we perform prediction and inference?
 - Lasso, but: convex relaxation; one parameter for sparsity and shrinkage
 - Point mass mixture prior. but: computation is prohibitive

Introduction		
Introduction		

Can we find a continuous prior that behaves like the point mass mixture prior?

Introduction		

Introduction

- Can we find a continuous prior that behaves like the point mass mixture prior?
- Desiderata:
 - adaptive to sparsity
 - easy to compute
 - good predictive performance
 - good frequentist properties
 - decent compromise between statistical and computational goals

Introduction		

Introduction

- Can we find a continuous prior that behaves like the point mass mixture prior?
- Desiderata:
 - adaptive to sparsity
 - easy to compute
 - good predictive performance
 - good frequentist properties
 - decent compromise between statistical and computational goals
- Global-local priors can achieve this (with some qualifications).

Introduction		

Introduction

- Can we find a continuous prior that behaves like the point mass mixture prior?
- Desiderata:
 - adaptive to sparsity
 - easy to compute
 - good predictive performance
 - good frequentist properties
 - decent compromise between statistical and computational goals
- Global-local priors can achieve this (with some qualifications).
- But... they are still slow.
 - Lasso: *n* ≈ 1,000, *p* ≈ 1,000,000;
 - Global-local: $n \approx 1,000, p \approx 1,000$.

Introduction	Model	Computation	Results	Conclusion
Model				
► The	Horseshoe model * :			
	$y_i \mid \beta_j$,	$\lambda_j, au, \sigma^2 \stackrel{ind}{\sim} N(x_ieta, \sigma^2)$		
		$eta_j \stackrel{\textit{ind}}{\sim} N(0, au^2 \lambda_j^2)$		
		$\lambda_{j} \stackrel{\textit{ind}}{\sim} Cauchy_{+}$ (0,	1)	
		$ au \sim {\sf Cauchy}_+$ (0, 1	1)	

 $\sigma^2 \sim \text{InvGamma}(a_0/2, b_0/2)$

*[Carvalho et. al, 2010]

	Model
Model	

► The Horseshoe model^{*}:

$$egin{aligned} y_i \mid eta_j, \lambda_j, au, \sigma^2 \stackrel{\textit{ind}}{\sim} N(x_ieta, \sigma^2) \ η_j \stackrel{\textit{ind}}{\sim} N(0, au^2\lambda_j^2) \ &\lambda_j \stackrel{\textit{ind}}{\sim} ext{Cauchy}_+(0, \ & au \sim ext{Cauchy}_+(0, \ & axt{Cauchy}_+(0, \ & axt$$

 $\sigma^2 \sim \text{InvGamma}(a_0/2, b_0/2)$

*[Carvalho et. al, 2010]

Introduction	Model	Computation	Results
Model			
► The	Horseshoe model [*] :	$\mathbf{x}_{i}, \mathbf{\tau}, \sigma^{2} \stackrel{ind}{\sim} \mathcal{N}(\mathbf{x}_{i} \mathbf{eta}, \sigma^{2})$	
	$y_i \mid p_j, \gamma$	$(j, 1, 0) \sim N(x_i p, 0)$	

 $\sigma^2 \sim \text{InvGamma}(a_0/2, b_0/2)$

*[Carvalho et. al, 2010]

Introduction	Model	Computation	Results	Conclusion
Madal				
Model				
► Th	ne Horseshoe model [*] :			
	$y_i \mid oldsymbol{eta}_j$, λ	$\sigma_{ij}, \tau, \sigma^2 \stackrel{ind}{\sim} N(x_i eta, \sigma^2)$		
		$eta_j \stackrel{\textit{ind}}{\sim} N(0, au^2 \lambda_j^2)$		

 $au \sim {
m Cauchy}_+(0,1)$

 $\sigma^2 \sim \text{InvGamma}(a_0/2, b_0/2)$

*[Carvalho et. al, 2010]

Introduction	Model	Computation	Results	Conclusion
Model				
► The	Horseshoe model * :			
	$y_i \mid \beta_j$,	$\lambda_j, au, \sigma^2 \stackrel{ind}{\sim} N(x_ieta, \sigma^2)$		
		$eta_j \stackrel{\textit{ind}}{\sim} N(0, au^2 \lambda_j^2)$		
		$\lambda_{j} \stackrel{\textit{ind}}{\sim} Cauchy_{+}$ (0,	1)	
		$ au \sim {\sf Cauchy}_+$ (0, 1	1)	

 $\sigma^2 \sim \text{InvGamma}(a_0/2, b_0/2)$

*[Carvalho et. al, 2010]

Model	
Model	

Horseshoe has other good frequentist properties.

	Model		Conclusion
Model			
Woder			

- Horseshoe has other good frequentist properties.
- ▶ It achieves the minimax-adaptive risk for squared error loss up to a constant.

	Model		
Model			
Model			

- Horseshoe has other good frequentist properties.
- It achieves the minimax-adaptive risk for squared error loss up to a constant.
- Suppose X = I, $\|\beta\|_0 = s_n$, then [van der Pas et al., 2014],

$$\sup_{\beta: \|\beta\|_0 \le s_n} \mathbb{E}_{\beta} \left[\|\hat{\beta}_{HS} - \beta\|_2^2 \right] \le 4\sigma^2 s_n \log \frac{n}{s_n} \cdot (1 + o(1)),$$

while, for any estimator $\hat{\beta}$, [Donoho et al., 1992] shows

$$\sup_{\beta: \|\beta\|_0 \leq s_n} \mathbb{E}_{\beta} \left[\|\hat{\beta} - \beta\|_2^2 \right] \geq 2\sigma^2 s_n \log \frac{n}{s_n} \cdot (1 + o(1)).$$

	Computation	Conclusion
Computation		

State-of-the-art: (i) $\tau \mid \beta, \sigma^2, \lambda$, (ii) $(\beta, \sigma^2) \mid \tau, \lambda$, (iii) slice sampling for λ .

	Computation	Conclusion
Computation		
Computation		

State-of-the-art: (i) $\tau | \beta, \sigma^2, \lambda$, (ii) $(\beta, \sigma^2) | \tau, \lambda$, (iii) slice sampling for λ . But...

	Computation	Conclusion
Computation		
computation		

- State-of-the-art: (i) $\tau | \beta, \sigma^2, \lambda$, (ii) $(\beta, \sigma^2) | \tau, \lambda$, (iii) slice sampling for λ . But...
- We scale the model with two ideas.

	Computation	
Computation		

- State-of-the-art: (i) $\tau | \beta, \sigma^2, \lambda$, (ii) $(\beta, \sigma^2) | \tau, \lambda$, (iii) slice sampling for λ . But...
- We scale the model with two ideas.
- First idea: **block** (β , σ^2 , τ) to improve *mixing*;
 - 1. sample $(\beta, \sigma^2, \tau) \mid \lambda$ by block sampling: $\tau \mid \lambda$, then $\sigma^2 \mid \tau, \lambda$, and finally $\beta \mid \sigma^2, \tau, \lambda$;
 - 2. sample $\lambda \mid \beta, \sigma^2$ using slice sampling.

	Computation	
Computation		

- State-of-the-art: (i) $\tau | \beta, \sigma^2, \lambda$, (ii) $(\beta, \sigma^2) | \tau, \lambda$, (iii) slice sampling for λ . But...
- We scale the model with two ideas.
- First idea: **block** (β , σ^2 , τ) to improve *mixing*;
 - 1. sample $(\beta, \sigma^2, \tau) \mid \lambda$ by block sampling: $\tau \mid \lambda$, then $\sigma^2 \mid \tau, \lambda$, and finally $\beta \mid \sigma^2, \tau, \lambda$;
 - 2. sample $\lambda \mid \beta, \sigma^2$ using slice sampling.
- Second idea: truncate some of the matrices involved to improve the *computational* cost per step.

Introduction	Model	Computation	Results	Conclusion
Gibbs sampling	I			
		T, $\xi = \tau^{-2}$, $\eta_j = \lambda_j^{-2}$, and $\frac{1}{2} \left(y^T M^{-1} y + b_0 \right)^{-\frac{n+a_0}{2}}$	block update:	

 $\blacktriangleright \ p(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif} \left[0, \frac{1}{1+\eta_j}\right]; \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\delta \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$

Introduction	Model	Computation	Results	Conclusion
Gibbs sampling	g			
		$\xi = au^{-2}, \ \eta_j = \lambda_j^{-2}, \ ext{and}$	block update:	

 $\blacktriangleright p(\beta \mid \sigma^2, \tau, \lambda, y) \sim N\left((X^T X + \operatorname{diag}(\xi\eta))^{-1} X^T y, \sigma^2 (X^T X + \operatorname{diag}(\xi\eta))^{-1} \right)$

Then perform slice sampling:

 $\blacktriangleright \ \rho(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif} \left[0, \frac{1}{1+\eta_j}\right]; \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\delta \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$

Gibbs sampling
Let
$$M = X(\operatorname{diag}(\xi\eta))^{-1}X^T + I$$
, $\xi = \tau^{-2}$, $\eta_j = \lambda_j^{-2}$, and **block update**:
 $p(\tau \mid \lambda, y) \propto \frac{1}{\sqrt{\xi(1+\xi)}} |M|^{-1/2} (y^T M^{-1}y + b_0)^{-\frac{n+a_0}{2}}$
 $p(\sigma^2 \mid \tau, \lambda, y) \sim \operatorname{InvGamma}\left(\frac{n+a_0}{2}, \frac{1}{2} [y^T M^{-1}y + b_0]\right)$
 $p(\beta \mid \sigma^2, \tau, \lambda, y) \sim N\left((X^T X + \operatorname{diag}(\xi\eta))^{-1}X^T y, \sigma^2(X^T X + \operatorname{diag}(\xi\eta))^{-1}\right)$

 $\blacktriangleright \ \rho(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif} \left[0, \frac{1}{1+\eta_j}\right]; \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\delta \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$

$$\begin{array}{l|c|c|c|c|c|} \hline \text{Model} & \hline \text{Computation} & \hline \text{Results} & \hline \text{Conclusion} \\ \hline \text{Gibbs sampling} \\ \hline \text{Gibbs sampling} \\ \hline \text{Let } M = X(\operatorname{diag}(\xi\eta))^{-1}X^T + I, \ \xi = \tau^{-2}, \ \eta_j = \lambda_j^{-2}, \ \text{and } \textbf{block update:} \\ \hline p(\tau \mid \lambda, y) \propto \frac{1}{\sqrt{\xi(1+\xi)}} |M|^{-1/2} \left(y^T M^{-1} y + b_0\right)^{-\frac{n+a_0}{2}} \\ \hline p(\sigma^2 \mid \tau, \lambda, y) \sim \operatorname{InvGamma}\left(\frac{n+a_0}{2}, \frac{1}{2}\left[y^T M^{-1} y + b_0\right]\right) \\ \hline p(\beta \mid \sigma^2, \tau, \lambda, y) \sim N\left((X^T X + \operatorname{diag}(\xi\eta))^{-1}X^T y, \sigma^2(X^T X + \operatorname{diag}(\xi\eta))^{-1}\right) \end{array}$$

 $\blacktriangleright p(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif}\left[0, \frac{1}{1+\eta_j}\right]: \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\delta \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$

$$\begin{array}{l|c|c|c|c|c|} \hline \text{Model} & \hline \text{Computation} & \hline \text{Results} & \hline \text{Conclusion} \\ \hline \text{Gibbs sampling} \\ \hline \text{Let } M = X(\operatorname{diag}(\xi\eta))^{-1}X^T + I, \ \xi = \tau^{-2}, \ \eta_j = \lambda_j^{-2}, \ \text{and } \textbf{block update:} \\ \hline p(\tau \mid \lambda, y) \propto \frac{1}{\sqrt{\xi}(1+\xi)} |M|^{-1/2} \left(y^T M^{-1} y + b_0\right)^{-\frac{n+a_0}{2}} \\ \hline p(\sigma^2 \mid \tau, \lambda, y) \sim \operatorname{InvGamma}\left(\frac{n+a_0}{2}, \frac{1}{2}\left[y^T M^{-1} y + b_0\right]\right) \\ \hline p(\beta \mid \sigma^2, \tau, \lambda, y) \sim N\left((X^T X + \operatorname{diag}(\xi\eta))^{-1}X^T y, \sigma^2(X^T X + \operatorname{diag}(\xi\eta))^{-1}\right) \end{array}$$

$$\blacktriangleright p(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif}\left[0, \frac{1}{1+\eta_j}\right]; \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\xi \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$$

IntroductionModelComputationResultsConclusionGibbs samplingLet
$$M = X(\operatorname{diag}(\xi\eta))^{-1}X^T + I$$
, $\xi = \tau^{-2}$, $\eta_j = \lambda_j^{-2}$, and block update: $p(\tau \mid \lambda, y) \propto \frac{1}{\sqrt{\xi}(1+\xi)} |M|^{-1/2} (y^T M^{-1}y + b_0)^{-\frac{n+a_0}{2}}$ $p(\sigma^2 \mid \tau, \lambda, y) \sim \operatorname{InvGamma}(\frac{n+a_0}{2}, \frac{1}{2} [y^T M^{-1}y + b_0])$ $p(\beta \mid \sigma^2, \tau, \lambda, y) \sim N((X^T X + \operatorname{diag}(\xi\eta))^{-1}X^T y, \sigma^2(X^T X + \operatorname{diag}(\xi\eta))^{-1})$

$$\blacktriangleright p(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif}\left[0, \frac{1}{1+\eta_j}\right]; \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\xi \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$$

Introduction Model Computation Results Conclusion
Gibbs sampling
Let
$$M = X(\operatorname{diag}(\xi\eta))^{-1}X^T + I$$
, $\xi = \tau^{-2}$, $\eta_j = \lambda_j^{-2}$, and block update:
 $p(\tau \mid \lambda, y) \propto \frac{1}{\sqrt{\xi(1+\xi)}} |M|^{-1/2} (y^T M^{-1}y + b_0)^{-\frac{n+a_0}{2}}$
 $p(\sigma^2 \mid \tau, \lambda, y) \sim \operatorname{InvGamma}\left(\frac{n+a_0}{2}, \frac{1}{2} [y^T M^{-1}y + b_0]\right)$
 $p(\beta \mid \sigma^2, \tau, \lambda, y) \sim N\left((X^T X + \operatorname{diag}(\xi\eta))^{-1}X^T y, \sigma^2(X^T X + \operatorname{diag}(\xi\eta))^{-1}\right)$

$$\blacktriangleright p(\lambda \mid \beta, \sigma^2, \tau, y): \text{ (i) } U \mid \eta_j \sim \text{Unif}\left[0, \frac{1}{1+\eta_j}\right]; \text{ (ii) } \eta_j \mid u \sim e^{-\frac{\xi \beta_j^2}{2\sigma^2}\eta_j} \mathbb{I}_{\left[\frac{1-u}{u} > \eta_j\right]}.$$

Introduction	Model	Computation	Results	Conclusion
Markov appro	ximation			
► We a	approximate $M = X$ diag(($(\xi\eta_j)^{-1})X^{ op}+1$ wit	h	
	$M_{\delta} = X D_{\delta} X^{T} +$	$I, \qquad D_{\delta} = diag($	$(\xi\eta_j)^{-1}\mathbb{I}_{[(\xi_{\max}\eta_j)^{-1}>\delta]})$	
for $\delta \ll 1$, and $\xi_{ m max}$ the maximum of the current and proposed $\xi.$				

Introduction	Model	Computation	Results	Conclusion
Markov approx	kimation			
► We a	oproximate $M = X$ diag(($\xi \eta$	$(y_j)^{-1})X^T + I$ wit	h	
	$M_{\delta} = X D_{\delta} X^{T} + I,$	$D_{\delta} = diag($	$(\xi\eta_j)^{-1}\mathbb{I}_{[(\xi_{\max}\eta_j)^{-1}>\delta]})$	
for $\delta \ll$ 1, and ξ_{\max} the maximum of the current and proposed ξ_{\cdot}				

Introduction	Model	Computation	Results	Conclusion
Markov appro	oximation			
► We a	approximate $M = X$ diag((8	$(\eta_j)^{-1})X^ op + I$ with	th	
	$M_{\delta} = X D_{\delta} X^{T} +$	$I, \qquad D_{\delta} = { m diag}($	$\mathbb{I}(\xi\eta_j)^{-1}\mathbb{I}_{[(\xi_{\max}\eta_j)^{-1}>\delta]})$	
for δ	$\xi \ll$ 1, and $\xi_{ m max}$ the maxim	um of the curren	t and proposed ξ .	
This	makes computation much	faster.		

Introduction	Model	Computation	Results	Conclusion
Markov appro	ximation			
► We a	pproximate $M = X$ diag(($\xi \eta$	$(j)^{-1})X^T + I$ with	th	
	$M_{\delta} = X D_{\delta} X^{T} + I,$	$D_{\delta} = diag($	$[(\xi\eta_j)^{-1}\mathbb{I}_{[(\xi_{\max}\eta_j)^{-1}>\delta]})$	
for δ	\ll 1, and $\xi_{\sf max}$ the maximum	m of the current	t and proposed ξ .	
► This	makes computation much f	aster.		
Approxim	ating Kernels			
Lot D.(Y) and $\mathcal{D}(\mathbf{x}_{1})$ denote the M	larkov operator	for the approximate and eva	ct

Let $\mathcal{P}_{\delta}(x, \cdot)$ and $\mathcal{P}(x, \cdot)$ denote the Markov operators for the approximate and exact algorithms, with $x = (\beta, \sigma^2, \tau, \lambda)$ the entire state vector. Then

$$\sup_{x} \|\mathcal{P}_{\delta}(x,\cdot) - \mathcal{P}(x,\cdot)\|_{\mathsf{TV}} \leq \sqrt{\delta} \|X\| \sqrt{a + \frac{n+a_0}{b_0} + \frac{n}{2} \frac{\|y\|^2}{b_0}} + \mathcal{O}(\delta),$$

for sufficiently small $\delta > 0$.

		Results	Conclusion
Simulation			

We simulate data as follows:

$$\begin{aligned} x_i &\stackrel{\text{iid}}{\sim} N_p(0, \boldsymbol{\Sigma}) \\ y_i &\sim N(x_i \beta, 4) \\ \beta_j &= \begin{cases} 2^{-(j/4-9/4)} & \text{if } j < 24, \\ 0 & \text{if } j \geq 24. \end{cases} \end{aligned}$$

		Results	Conclusion
Simulation			
Sinuation			

We simulate data as follows:

$$\begin{split} x_i &\stackrel{\text{id}}{\to} N_p(0, \mathbf{\Sigma}) \\ y_i &\sim N(x_i\beta, 4) \\ \beta_j &= \begin{cases} 2^{-(j/4-9/4)} & \text{if } j < 24, \\ 0 & \text{if } j \geq 24. \end{cases} \end{split}$$

▶ There are nulls, clear non-nulls, and some subtle non-nulls.

Introduction	Model	Computation	Results	Conclusion
Simulation				
Simulation				

We simulate data as follows:

$$\begin{split} x_i &\stackrel{\text{id}}{\to} N_p(0, \mathbf{\Sigma}) \\ y_i &\sim N(x_i \beta, 4) \\ \beta_j &= \begin{cases} 2^{-(j/4-9/4)} & \text{if } j < 24, \\ 0 & \text{if } j \geq 24. \end{cases} \end{split}$$

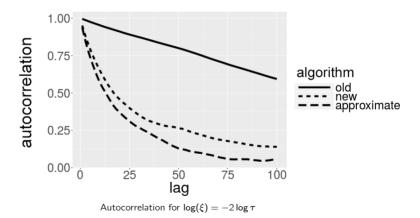
▶ There are nulls, clear non-nulls, and some subtle non-nulls.

• We consider both $\Sigma = I$ (independent design) and $\Sigma_{ij} = 0.9^{|i-j|}$ (correlated design).

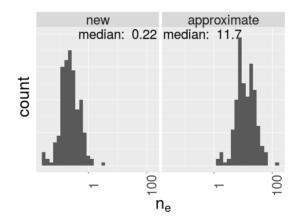
Model

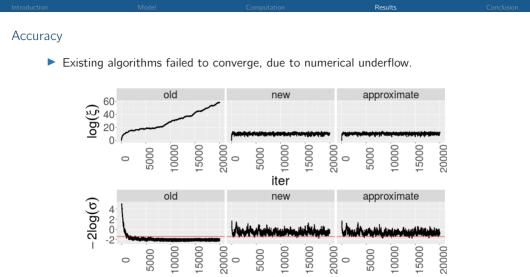
Computation

Autocorrelation



Approximate algorithm is $50 \times$ more efficient with n = 2,000 and p = 20,000.



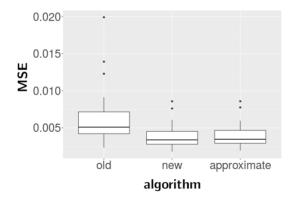


Trace plots for $-2\log(\sigma)$ and $\log(\xi) = -2\log(\tau)$; truth in red

iter

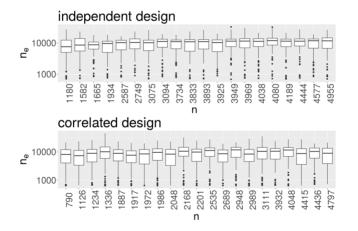
Accuracy

▶ In terms of MSE, the approximation costs us little.



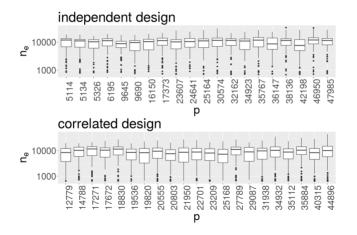
Dependence on p and n

Effective sample sizes seem independent of n and p.



Dependence on *p* and *n*

Effective sample sizes seem independent of n and p.



Real application: GWAS

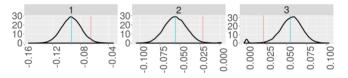
> n = 2267 observations, p = 98385 SNPs in the genome of maize.

Real application: GWAS

- ▶ n = 2267 observations, p = 98385 SNPs in the genome of maize.
- X: maize seeds; y: growing degree days to silking ('growth cycle')

 \triangleright n = 2267 observations, p = 98385 SNPs in the genome of maize.

 \triangleright X: maize seeds; v: growing degree days to silking ('growth cycle')

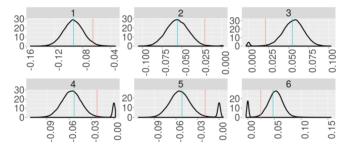


Bimodal posterior distribution for $\beta \mid v$: Lasso (red) shrinks more than Horseshoe (blue)

Paulo Orenstein

▶ n = 2267 observations, p = 98385 SNPs in the genome of maize.

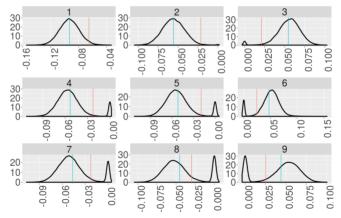
X: maize seeds; y: growing degree days to silking ('growth cycle')



Bimodal posterior distribution for $\beta \mid y$; Lasso (red) shrinks more than Horseshoe (blue)

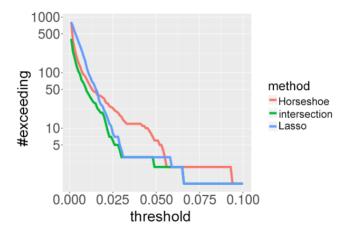
▶ n = 2267 observations, p = 98385 SNPs in the genome of maize.

X: maize seeds; y: growing degree days to silking ('growth cycle')



Bimodal posterior distribution for $\beta \mid y$; Lasso (red) shrinks more than Horseshoe (blue)

Variable selection with Horseshoe



Number of variables for which $\hat{\beta}_{HS,j} = \mathbb{E}[\beta_j \mid y] > t$ or $\hat{\beta}_{Lasso,j} > t$ vs threshold t; both methods largely agree on the identities of the signals

		Conclusion
Conclusion		

There is no point in having a great model, like the Horseshoe, if it can't be computed.

		Conclusion
Conclusion		

- There is no point in having a great model, like the Horseshoe, if it can't be computed.
- ▶ There is a need to scale more Bayesian models to the level of Frequentists.

		Conclusion
Conclusion		

- There is no point in having a great model, like the Horseshoe, if it can't be computed.
- ▶ There is a need to scale more Bayesian models to the level of Frequentists.
- We manage to do that for the Horseshoe prior with two ideas: blocking and truncation.

		Conclusion
Conclusion		

- There is no point in having a great model, like the Horseshoe, if it can't be computed.
- ▶ There is a need to scale more Bayesian models to the level of Frequentists.
- We manage to do that for the Horseshoe prior with two ideas: blocking and truncation.
- We observed interesting and novel statistical phenomena, e.g., bimodality of β .

		Conclusion
Conclusion		

- There is no point in having a great model, like the Horseshoe, if it can't be computed.
- ▶ There is a need to scale more Bayesian models to the level of Frequentists.
- We manage to do that for the Horseshoe prior with two ideas: blocking and truncation.
- We observed interesting and novel statistical phenomena, e.g., bimodality of β .
- ▶ There is likely more room for improvement.

Introduction	Model	Computation	Results	Conclusion
References				
 Bhattacharya (2012). 	a, Anirban, et al.	"Bayesian shrinkage".	arXiv preprint, arXiv: .	1212.6088
		as G. Polson, and Jai <i>Biometrika</i> 97.2 (201		horseshoe

- Johndrow, James E., and Jonathan C. Mattingly. "Error bounds for Approximations of Markov chains." arXiv preprint, arXiv:1711.05382 (2017).
- Johndrow, James E., P. O., Bhattacharya, Anirban "Scalable MCMC for Bayes Shrinkage Priors". arXiv preprint, arXiv: 1705.00841 (2018).
- Rudolf, Daniel, and Nikolaus Schweizer. "Perturbation theory for Markov chains via Wasserstein distance." *Bernoulli* 24.4A (2018): 2610-2639.
- Van Der Pas, S. L., B. J. K. Kleijn, and A. W. Van Der Vaart. "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal* of *Statistics* 8.2 (2014): 2585-2618.

Mode

Computation

Conclusion

Extra slides

More simulation results

► Why "Horseshoe"?

Introduction	Model	Computation	Results	Conclusion
More simulation	ns			
We let	n = 1000 and $p = 2$	0,000.		
4				
3				



2

1

0

Scalable MCMC for Bayes Shrinkage Priors

peta

More simulations
4.
3.
2.
peta 5444444444444444444444444444444444444

Paulo Orenstein

More simulations
3·
2.
¹੶ ੶੶ ੶੶੶ ₀੶੶੶੶::::::::::::::::::::::::::
peta ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Paulo Orenstein

Introduction	Model	Computation	Results	C
More simulations				
4				
3.				
2.				
1· •• 🛉	•			



beta

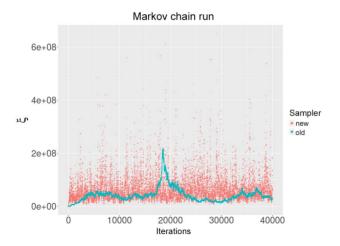
Paulo Orenstein

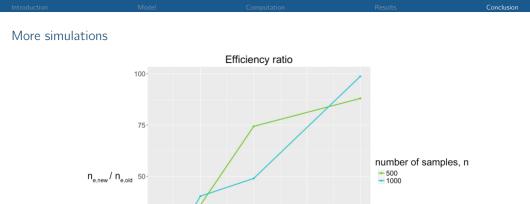
Conclusion

		Conclusion

More simulations

▶ The new algorithm lead to significant improvement in the autocorrelation:





25-

5000

number of features, p

15000

20000

4

		Conclusion

Why "Horseshoe"?

In the orthogonal case with $n \ge p$ and $\sigma^2 = \tau = 1$, and defining a shrinkage profile $\kappa_j = 1/(1 + n\lambda_j^2)$, we can write $\mathbb{E}[\beta_j|y] = (1 - \mathbb{E}[\kappa_j|y])\hat{\beta}_j$.

2

		Conclusion

Why "Horseshoe"?

In the orthogonal case with $n \ge p$ and $\sigma^2 = \tau = 1$, and defining a shrinkage profile $\kappa_j = 1/(1 + n\lambda_j^2)$, we can write $\mathbb{E}[\beta_j|y] = (1 - \mathbb{E}[\kappa_j|y])\hat{\beta}_j$.

2

		Conclusion

Why "Horseshoe"?

- In the orthogonal case with $n \ge p$ and $\sigma^2 = \tau = 1$, and defining a shrinkage profile $\kappa_j = 1/(1 + n\lambda_j^2)$, we can write $\mathbb{E}[\beta_j|y] = (1 \mathbb{E}[\kappa_j|y])\hat{\beta}_j$.
- Prior density for κ_i :

