Blocked Gibbs Sampler with Anti-correlation Gaussian Data Augmentation, with Applications to L1-ball Prior Models

Yu Zheng and Leo L. Duan

Department of Statistics University of Florida



1 Introduction

- Problem Setting & Algorithm
- 3 Numerical Experiments
- One More Application

5 Discussion



Introduction

- There is a rich literature on Bayesian sparse modeling.
- Applications include variable selection in regression with p ≫ n, piece-wise constant smoothing where most of contrasts between parameters are zero, etc.
- We focus on the MCMC sampling algorithm for variable selection problems, under augmented likelihood with θ in a Gaussian density

$$L(y; \theta, z) \propto \exp(-\frac{1}{2}\theta' M_z \theta + \theta' m_z).$$

• We focus on exactly sparse models, with $\theta_j = 0$ happens with > 0 probability (in both prior and posterior).



Commonly Used Exactly Sparse Priors

• Discrete spike-and-slab priors (Mitchell and Beauchamp (1988)):

$$\Pi_{0}(\theta) = \prod_{j=1}^{p} \left[w \underbrace{\delta_{0}(\theta_{j})}_{\text{spike}} + (1-w) \underbrace{f(\theta_{j})}_{\text{slab}} \right]$$

• *I*₁-ball priors (Xu and Duan 2023):

$$eta \sim \Pi_0^eta, \qquad heta = \operatorname{argmin}_{z: \|z\|_1 \leq r} \|z - eta\|_2^2$$

or, with reparameterization:

$$eta \sim \mathsf{\Pi}^eta_{\mathsf{0}}, \qquad heta = \mathsf{sign}(eta) \circ (|eta| - \kappa)_+$$

< □ > < 同 > < 回 > < 回 < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < < □ < □ < < □ < □ < < □ < < □ < □ < < □ < □ < □ < < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ < □ <

- For a Gaussian linear regression using spike-and-slab priors, very efficient algorithms have been developed since one can marginalize most of the parameters, leading to collapsed Gibbs sampler:
 - Orthogonal Data Augmentation (ODA) (Ghosh and Clyde (2011))
 - Stochastic Search Variable Selection (SSVS) (George and McCulloch (1995))
- For design matrix that is high-dimensional or contains highly correlated predictors:
 - Shotgun algorithm (Hans et al. (2007))
 - Parallel tempering (Bottolo and Richardson (2010))
 - Correlation-based search (Kwon et al. (2011))
 - Two-parameter flipping Metropolis-Hastings algorithm under g-prior slab (Yang et al. (2016))

イロト イヨト イヨト

I₁-Ball Priors

- The spike-and-slab priors assume the independence for different entries of θ a priori: $\Pi_{0}(\theta) = \prod_{j=1}^{p} \left[w \delta_{0}(\theta_{j}) + (1-w) f(\theta_{j}) \right]$
- Recent interest in "structured sparsity" in the sense that:
 - The occurrences of zeros (or close-to-zero) are dependent, according to a temporal, spatial, or group structure
 - 2 The non-zeros could also be correlated
- I_1 -ball priors are quite convenient for addressing such modeling needs:

$$\beta \sim \Pi_0^{\beta}, \qquad \theta = \operatorname{sign}(\beta) \circ (|\beta| - \kappa)_+$$

- One could allow β to have some dependence structure, such as from a Gaussian process (Kang et al. (2018))
- One could let κ to be a vector with values following a pre-defined group structure (Yuan and Lin (2006); Bai et al. (2022))

- The soft-thresholding transform has two useful properties: continuity and smoothness almost everywhere (w.r.t. Π_0^{β}).
- It allows for the gradient-based hybrid Monte Carlo algorithms such as No-U-Turn Hamiltonian Monte Carlo.
- HMC has caveats of being sensitive to tuning and may have a high computational cost due to the gradient evaluation.
- A simpler alternative, the Gibbs sampler, is tuning-free and hence quite friendly to general users.

Let $\theta \in \mathbb{R}^{p}$ be the parameter of interest. We focus on the model with conditional posterior taking the following form

$$\Pi(\theta,\beta \mid M,\phi,H,\psi) \propto \exp\{-\frac{1}{2}(\theta'M\theta - 2\phi'\theta)\}\exp\{-\frac{1}{2}(\beta'H\beta - 2\psi'\beta)\},\$$

$$\theta = \operatorname{sign}(\beta) \circ (|\beta| - \kappa)_{+}.$$
 (1)

Two examples:

- Variable selection in a linear regression model: $y = X\theta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \Omega^{-1})$, $\beta_j \stackrel{indep}{\sim} \mathcal{N}(0, \tau_j) \rightarrow M = X'\Omega X$, $\phi = X'\Omega y$, $H = \text{diag}(1/\tau_j)$ and $\psi = 0$.
- Sparse smoothing model: $y = \theta + \epsilon$, with y_i associated with some spatial coordinate s_i , $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and $\beta \sim \mathcal{N}(0, K(s, s)) \rightarrow M = \sigma^{-2}I$, $\phi = \sigma^{-2}y$, $H = K^{-1}(s, s)$ and $\psi = 0$.

- For any non-diagonal M or H, the quadratic terms $\theta' M \theta$ and $\beta' H \beta$ make it difficult to explore a significant change in the parameters.
- The correlation within entries of θ or β (a posteriori) adds a computational burden to the posterior sampling using a Gibbs sampler.

Q: Can we cancel out the quadratic terms $\theta' M \theta$ and $\beta' H \beta$, with some clever data augmentation trick?



Consider latent variables $r, t \in \mathbb{R}^p$:

$$(r \mid \theta, M) \sim \mathcal{N} \{ (dI_p - M)\theta, \ (dI_p - M) \}, (t \mid \beta, H) \sim \mathcal{N} \{ (eI_p - H)\beta, \ (eI_p - H) \},$$
(2)

where d, e > 0 are two constants chosen to make $dI_p - M$ and $eI_p - H$ positive definite.

- One can take d to be "slightly" greater than the largest eigenvalue of M, say $\lambda_p(M) + 10^{-4}$.
- We refer to (2) as the "anti-correlation Gaussian" in that they cancel out the correlation between different entries of θ/β (We will see this soon).

Combining the posterior (1) of (θ, β) and the conditional distribution (2) of (r, t), we have the distribution of (θ, β) given (r, t):

$$\Pi(\theta,\beta|r,t,M,H,\phi,\psi)$$

$$\times \Pi(\theta,\beta|N,H,\phi,\psi) \Pi(r,t|\theta,\beta,M,H,\phi,\psi)$$

$$\times \underbrace{\exp\{-\frac{1}{2}(\theta'M\theta-2\phi'\theta)\}\exp\{-\frac{1}{2}(\beta'H\beta-2\psi'\beta)\}}_{\Pi(\theta,\beta|M,H,\phi,\psi)}$$

$$\cdot \underbrace{\exp\{-\frac{1}{2}(\theta'(dI_p-M)\theta-2r'\theta+\beta'(eI_p-H)\beta-2t'\beta)\}}_{\Pi(r,t|\theta,\beta,M,H,\phi,\psi)}$$

イロト イヨト イヨト

Data Augmentation

$$\begin{split} & \Pi(\theta,\beta|r,t,M,H,\phi,\psi) \\ \propto \Pi(\theta,\beta|N,H,\phi,\psi) \Pi(r,t|\theta,\beta,M,H,\phi,\psi) \\ \propto \underbrace{\exp\{-\frac{1}{2}(\theta'M\theta-2\phi'\theta)\}\exp\{-\frac{1}{2}(\beta'H\beta-2\psi'\beta)\}}_{\Pi(\theta,\beta|M,H,\phi,\psi)} \\ & \cdot \underbrace{\exp\{-\frac{1}{2}(\theta'(dI_p-\mathcal{M})\theta-2r'\theta+\beta'(eI_p-\mathcal{M})\beta-2t'\beta)\}}_{\Pi(r,t|\theta,\beta,M,H,\phi,\psi)} \\ = \prod_{j=1}^{p} \exp\left\{-\frac{1}{2}\left[d\theta_j^2-2(\phi_j+r_j)\theta_j+e\beta_j^2-2(\psi_j+t_j)\beta_j\right]\right\}, \\ & \theta_j = \operatorname{sign}(\beta_j)(|\beta_j|-\kappa_j)_+. \end{split}$$

 $(heta_j,eta_j)$'s are now conditionally independent!

Yu Zheng (UF)

UF FLORIDA

э

イロト 不得 トイヨト イヨト

$$\Pi(\theta,\beta|r,t) \propto \prod_{j=1}^{p} \exp\left\{-\frac{1}{2}\left[d\theta_{j}^{2}-2(\phi_{j}+r_{j})\theta_{j}+e\beta_{j}^{2}-2(\psi_{j}+t_{j})\beta_{j}\right]\right\},\$$

$$\theta_{j}=\operatorname{sign}(\beta_{j})(|\beta_{j}|-\kappa_{j})_{+}.$$

- The conditional independence over j allows us to draw β_j 's in a block.
- β_j follows a three-component mixture and can be drawn in two steps:
 - Oraw a discrete variable b_j = sign(θ_j) ∈ {-1, 0, 1}
 Oraw β_j according to b_j:

$$(\beta_{j} \mid b_{j} = 0) \sim \mathcal{N}_{(-\kappa_{j},\kappa_{j})}(\frac{\psi_{j} + t_{j}}{e}, \frac{1}{e}),$$

$$(\beta_{j} \mid b_{j} = 1) \sim \mathcal{N}_{(\kappa_{j},\infty)}(\frac{\phi_{j} + r_{j} + \psi_{j} + t_{j} + d\kappa_{j}}{d + e}, \frac{1}{d + e}),$$

$$(\beta_{j} \mid b_{j} = -1) \sim \mathcal{N}_{(-\infty,-\kappa_{j})}(\frac{\phi_{j} + r_{j} + \psi_{j} + t_{j} - d\kappa_{j}}{d + e}, \frac{1}{d + e}).$$

$$(3)$$

▶ ∢ ⊒ ▶

Numerical Experiments

• I_1 -ball Variable selection:

$$\begin{split} y_i &\sim \mathcal{N}(x_i'\theta, \sigma^2), i = 1, \cdots, n, \\ \sigma^2 &\sim \mathsf{InvGamma}(a_\sigma, b_\sigma), \\ \theta_j &= \mathsf{sign}(\beta_j)(|\beta_j| - \kappa_j)_+, j = 1, \cdots, p, \\ \beta_j &\stackrel{indep}{\sim} \mathcal{N}(0, \tau_j), \\ \tau_j &\sim \mathsf{InvGamma}(a_j, b_j), \\ \kappa_j &\sim \mathsf{Exp}(\lambda_j). \end{split}$$

The ground truth: the first $d_0 = 10$ entries of θ are from $\mathcal{N}(5, 0.5)$, and the rest is set to be zero.

• One can also assume a common threshold $\kappa \sim \text{Exp}(\lambda)$.

イロト イヨト イヨト イヨ

Results



590 15 / 26

Results



Yu Zheng (UF)

Yu Zheng (UF)

Sparse smoothing model: y = θ + ε, where ε ~ No(0, Iσ²) and θ has spatially correlated entries θ_i, each of which corresponds to a spatial coordinate s_i. We have spatially correlated Gaussian precursor β from No{0, K(s, s)} with K Gaussian kernel and θ is obtained by thresholding β with a communal threshold κ₀.

Results

theta (ground true) y (non-smooth) theta (posterior mean) 1.0 1.0 1.0 0.8 0.8 0.8 0.6 0.6 0.6 0.4 0.4 0.4 0.2 0.2 0.2 0.0 0.0 0.0 0.2 0.8 0.0 0.4 0.8 0.0 0.4 0.6 1.0 0.2 0.4 0.6 0.8 1.0 0.0 0.2 0.6 1.0

Figure: Sparse smoothing

VERSITY of ORIDA

Ē

- Although there has been a great number of algorithms designed for the spike-and-slab priors that enjoy both high efficiency and fast mixing in numerical experiments, the convergence properties of the Markov chain have not yet been discussed enough.
- For simplicity, the remaining discussion assumes $(M, \phi, H, \psi, \kappa)$ are known and remain unchanged.

Let $\Phi_{BGS} = \{(\theta_m, \beta_m, r_m, t_m)\}_{m=0}^{\infty}$ denote the Markov chain generated by our Blocked Gibbs sampler. It has an invariant distribution $w(\theta, \beta, r, t)$.

- Suppose we want to estimate $\mathbb{E}_w g(\theta, \beta, r, t)$ for some real-valued measurable function g.
- A strongly consistent estimator is $\bar{g}_m = \frac{1}{m+1} \sum_{k=0}^m g(\theta_k, \beta_k, r_k, t_k)$ for a Harris ergodic Markov chain.
- Establishing the geometric ergodicity leads to a Markov chain Central Limit Theorem (CLT):

$$\sqrt{m}(\bar{g}_m - \mathbb{E}_{\Pi_\infty}g) o \mathcal{N}(0, \sigma_g^2).$$

• To get a consistent estimator of the standard error for \bar{g}_m , we require such CLT to hold.

One way to establish the geometric ergodicity is through proving a drift condition and a minorization condition for (r, t)-block (Theorem 12 in Rosenthal (1994); Meyn and Tweedie (1994)):

(Drift condition): There exist some $V : \mathcal{X} \to \mathbb{R}^{\geq 0}$, $\lambda < 1$, and $b < \infty$ such that:

$$\mathbb{E}[V(r_{m+1},t_{m+1})|r_m,t_m] \leq \lambda V(r_m,t_m) + b.$$

(Minorization condition): There exist some $\epsilon > 0$, some probability measure $Q(\cdot)$ on \mathcal{X} , and $d > 2b/(1-\lambda)$ such that for $\forall (r_0, t_0) \in \mathcal{X}$ with $V(r_0, t_0) < d$,

 $P_{r,t}((r_0, t_0), \cdot) \geq \epsilon Q(\cdot).$

(日) (四) (日) (日) (日)

Theorem

The Markov chain Φ_{BGS} is geometrically ergodic. That is, there exist a real-valued function $C(\theta, \beta, r, t)$ and $0 < \gamma < 1$ such that for all $(\theta_0, \beta_0, r_0, t_0) \in \mathcal{X} \times \mathcal{Y}$,

$$\|P^m_{BGS}((\theta_0,\beta_0,r_0,t_0),\cdot)-w(\cdot)\|_{TV}\leq C(\theta_0,\beta_0,r_0,t_0)\gamma^m.$$

More details can be found in our pre-print.

- We proved the geometric ergodicity for our Markov chain (the Gibbs sampler with anti-correlation Gaussian data augmentation).
- To the best of our knowledge, this is the first work on the proof of geometric ergodicity for a Gibbs sampler with the spike-and-slab-type priors.

Extension: Sampling From Truncated Multivariate Normal

Suppose we wish to sample from a truncated multivariate normal:

$$heta \sim \mathcal{N}(\mu, \Sigma)$$
 subject to $heta_j \in R_j \subset \mathbb{R}, j = 1, \cdots, p.$

where R_j is some interval whose endpoints can be either $-\infty$ or ∞ . We introduce a latent variable r:

$$r| heta \sim \mathcal{N}((dI - \Sigma^{-1}) heta, dI - \Sigma^{-1}),$$

where d > 0 makes the matrix $dI - \Sigma^{-1}$ positive definite. Conditioned on the latent variable, the correlation between different entries of θ is canceled out, thus leading to

$$heta_j | r_j \stackrel{\textit{indep}}{\sim} \mathcal{N}_{\mathcal{R}_j}\left(rac{r_j + \phi_j}{d}, rac{1}{d}
ight),$$

where $\phi = \Sigma^{-1} \mu$.

Yu Zheng (UF)

- Our algorithm can be applied to many important scenarios spanning from variable selection to sparse smoothing.
- The novel data augmentation method, along with an efficient sampling algorithm of the latent variables, speeds up the Gibbs sampling and reaches a high effective sample size per second.
- The geometric ergodicity proof serves as an important theoretical guarantee for the convergence behavior of the Gibbs sampler under *l*₁-ball priors and justifies the broad usage of such modeling.
- Code and paper will be available soon on Github and arXiv.

Reference I

- Bai, R., G. E. Moran, J. L. Antonelli, Y. Chen, and M. R. Boland (2022). Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models. *Journal of the American Statistical Association 117*(537), 184–197.
- Bottolo, L. and S. Richardson (2010). Evolutionary Stochastic Search for Bayesian Model Exploration. *Bayesian Analysis* 5(3), 583–618.
- George, E. I. and R. E. McCulloch (1995). Stochastic Search Variable Selection. *Markov Chain Monte Carlo in Practice 68*, 203–214.
- Ghosh, J. and M. A. Clyde (2011). Rao–Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augmentation Approach. *Journal of the American Statistical Association 106*(495), 1041–1052.
- Hans, C., A. Dobra, and M. West (2007). Shotgun Stochastic Search for "Large P" Regression. Journal of the American Statistical Association 102(478), 507–516.
- Kang, J., B. J. Reich, and A.-M. Staicu (2018). Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process. *Biometrika* 105(1), 165–184.
- Kwon, D., M. T. Landi, M. Vannucci, H. J. Issaq, D. Prieto, and R. M. Pfeiffer (2011). An Efficient Stochastic Search for Bayesian Variable Selection With High-Dimensional Correlated Predictors. *Computational Statistics & Data Analysis* 55(10), 2807–2818.

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 - のへで

- Meyn, S. P. and R. L. Tweedie (1994). Computable Bounds for Geometric Convergence Rates of Markov Chains. *The Annals of Applied Probability* 4(4), 981–1011.
- Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association 83*(404), 1023–1032.
- Rosenthal, J. S. (1994). Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association 90*, 558–566.
- Yang, Y., M. J. Wainwright, and M. I. Jordan (2016). On the Computational Complexity of High-Dimensional Bayesian Variable Selection. *The Annals of Statistics* 44(6), 2497–2532.
- Yuan, M. and Y. Lin (2006). Model Selection and Estimation in Regression With Grouped Variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67.

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 - のへで