# **Motivating Problem**

In Bayesian sparse modeling, the spike-and-slab priors assume the independence for different entries of  $\theta$  a priori:  $\Pi_0(\theta) = \prod_{j=1}^p |w\delta_0(\theta_j) + (1-w)f(\theta_j)|$ . However, there is recent interest in "structured" sparsity in the sense that the occurrences of both zeros and non-zeros could be dependent.  $l_1$ -ball priors[2] are quite convenient for addressing such modeling needs:  $\beta \sim \Pi_0^{\beta}, \ \theta = \operatorname{sign}(\beta) \circ (|\beta| - \kappa)_+$ , where  $\theta \in \mathbb{R}^p$  is the parameter of interest, associated with precursor random variable  $\beta \in \mathbb{R}^p$ . We focus on the following posterior distribution:

$$\Pi(\theta, \beta \mid M, \phi, H, \psi, \kappa, \mathcal{Y}) \propto \exp\left[-\frac{1}{2}(\theta' M \theta - 2\phi' \theta + \beta' H \beta - 2\psi' \beta)\right]$$
$$\theta = \operatorname{sign}(\beta) \circ (|\beta| - \kappa)_{+},$$

For any non-diagonal M or H, the quadratic term  $\theta' M \theta$  or  $\beta' H \beta$  in the exponent of posterior makes it difficult to explore a large change in the parameter. This motivates us to use some latent variables to cancel out those terms.

## Anti-Correlation Gaussian Data Augmentation

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

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

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

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

 $(\theta_i, \beta_i)$ 's are now conditionally independent! The conditional independence over j allows us to draw  $\beta_i$ 's in a block.

# Efficient Sampling of the Anti-Correlation Gaussian in Regression

Sample  $r \sim \mathcal{N}[(dI - X'\Omega X)\theta, (dI - X'\Omega X)]$  efficiently when  $\Omega$  is complicated: We pre-compute the singular value decomposition (SVD) of  $X = U_X \Lambda_X V'_X$ .

- 1. Sample  $\gamma_1 \sim \mathcal{N}(0, dI_n)$ ,  $\gamma_2 \sim \mathcal{N}(0, dI_{p-n})$ ,  $\gamma_3 \sim \mathcal{N}[\Lambda_X \gamma_1/d, b_\Omega I_n (\Lambda_X)^2/d]$ ;
- 2. Sample  $\eta \sim \mathcal{N}(0, \Omega^{-1} b_{\Omega}I_n)$ ;

3. Set 
$$r = V_X \gamma_1 + V_X^{\dagger} \gamma_2 - X' \Omega (U_X \gamma_3 + \eta) + (dI - X' \Omega X) \theta$$
.

# GIBBS SAMPLING USING ANTI-CORRELATION GAUSSIAN DATA AUGMENTATION, WITH APPLICATIONS TO L1-BALL-TYPE MODELS UTT UNIVERSITY of FICTOR Yu Zheng<sup>1</sup> and Leo L. Duan<sup>1</sup> <sup>1</sup>Department of Statistics, University of Florida

# Extension: Sampling from Truncated MVN

One of the extensions is on the sampling of truncated multivariate Gaussian:

$$\Pi(\theta \mid \mu, \Sigma, R) \propto \exp\left[-\frac{1}{2}(\theta - \mu)^{\mathrm{T}}\Sigma^{-1}(\theta - \mu)\right]$$

where  $\mu \in \mathbb{R}^p$ ,  $\Sigma$  is positive definite, and R some constrained set of dimension p. Using the anti-correlation Gaussian  $(r \mid \theta, \mu, \Sigma) \sim \mathcal{N}[(dI - \Sigma^{-1})(\theta - \mu), dI - \Sigma^{-1}]$ , we have

$$\Pi(\theta \mid \mu, r) \propto 1(\theta \in R) \prod_{j=1}^{p} \exp\left[-\frac{1}{2}d(\theta_j - \mu_j)^2 + (\theta_j - \mu_j)^2\right]$$

When the constraints in R are separable over each sub-dimension, then  $\theta_i$  is conditionally independent over j.

# **Geometric Ergodicity**

### Consider the following Markov transition kernel:

 $\mathcal{K}(\theta^{m+1}, \beta^{m+1}, r^{m+1}, t^{m+1} \mid r^m, t^m)$  $= \prod_{\mathcal{K}} (r^{m+1} \mid \theta^{m+1}) \prod_{\mathcal{K}} (t^{m+1} \mid \beta^{m+1}) \prod_{\mathcal{K}} (\beta^{m+1}, \theta^{m+1} \mid r^m, t^m).$ 

**Theorem 1.** The Markov chain generated by  $\mathcal{K}(\theta^{m+1}, \beta^{m+1}, r^{m+1}, t^{m+1} | r^m, t^m)$  is geometrically ergodic. Specifically, there exists a real-valued function  $C_2(r^0, t^0)$  and  $0 < \gamma < 1$  such that for all  $(r^0, t^0)$ ,

 $\|P^{m}_{(\beta,r,t)}[(r^{0},t^{0}),\cdot] - \mu_{(\beta,r,t)}(\cdot)\|_{TV} \le C_{2}(r^{0},t^{0})\gamma^{m},$ 

where  $\|\cdot\|_{TV}$  denotes the total variation norm.

# **Simulation: Variable Selection**

Linear regression with  $y_i \sim N(x_i^{T}\theta, \sigma^2), x_i \in \mathbb{R}^p$  simulated from a multivariate Gaussian with mean zero, and correlation  $\rho^{|j-j'|}$  between  $x_{i,j}$  and  $x_{i,j'}$ . We set the first 10 elements of  $\theta$ to  $c\sqrt{\frac{\sigma^2 \log(p)}{n}}(2, -3, 2, 2, -3, 3, -2, 3, -2, 3)^T$ , where c is the selected signal-to-noise ratio, taken from  $\{1, 2, 3, 6\}$ . The other elements of  $\theta$  are set to zero.

p	Anti-corr Gaussian	NUTS	Comp-wise slice	EPC slice
10	0.41	26.18	3.61	3.04
50	0.93	108.33	31.44	4.16
500	5.60	22875.74	1121.47	16.14

Table 1: Running time for 1,000 iterations for the four algorithms. The time unit is in seconds based on pure R implementation for each algorithm.

$(p, \rho)$	Anti-corr Gaussian	NUTS	Comp-wise slice	EPC slice
(10, 0.5)	(202.36, 265,56)	(8.82, 7.05)	(49.08, 60.31)	(5.19, 8.23)
(50, 0.5)	(62.86, 165.74)	(1.52, 2.46)	(5.12, 4.61)	(4.34, 5.98)
(500, 0.5)	(4.81, 34.95)	(0.01, 0.01)	(0.10, 0.14)	(2.36, 7.58)
(10, 0.9)	(31.19, 35.79)	(4.22, 3.89)	No convergence	(3.61, 4.48)
(50, 0.9)	(11.28, 15.18)	(0.53, 1.05)	No convergence	No convergence
(500, 0.9)	(3.05, 14.50)	(<0.01, <0.01)	No convergence	No convergence

Table 2: Effective sample size per computing time (ESS/s) for the four algorithms. In each cell, the first number is the average ESS/s for the first 10 entries, and the second number is the average ESS/s for the rest entries.



 $1(\theta \in R)$ 

 $\theta_j - \mu_j r_j ].$ 

# Simulation: Soft-Thresholded Gaussian Process[1]

We consider image smoothing, in the form of:  $y_s = \theta_s + \epsilon_s$ ,  $\epsilon_s \stackrel{iid}{\sim}$  $\mathcal{N}(0,\sigma^2)$ , where s is the pixel location, s = (i,j) with  $i = 1, \ldots, n_1$ and  $j = 1, \ldots, n_2$ . We set the covariance function as K(s, s') = $\tau \exp[-\|s-s'\|_2^2/(2\xi^2)].$ 









truth  $\theta$ .

duced anti-correlation

# **Data Application**

In application, we use the functional magnetic resonance imaging scan of one human subject who was performing a motor task. We take a sectional view along the anterior-posterior axis of the brain, corresponding to  $91 \times 91$  pixels per frame, and over a time period, we collect 280 frames.

Using Anti-correlation Gaussian data augmentation, the effective sample size per hour for four randomly selected locations of  $\theta^{100}$  at f = 100: 315.14 at s = 3300, 315.44 at s = 3320, 268.95 at s = 3720, and 227.16 at s = 6700. Using NUTS, the effective sample size per hour for estimating  $\theta^{100}$  is 8.1 on average, 9.2 at s = 3300, 9.3 at s = 3320, 7.4 at s = 3720, and 6.7 at s = 6700.



# References

- [1] Jian Kang, Brian J Reich, and Ana-Maria Staicu. "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process". In: Biometrika 105.1 (2018), pp. 165–184.
- [2] Maoran Xu and Leo L Duan. "Bayesian Inference With the L1-Ball Prior: Solving Combinatorial Problems With Exact Zeros". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) in press (2023).



