

Bayesian High-dimensional Linear Regression with Sparse Projection-Posterior

Samhita Pal, Subhashis Ghosal

ICSDS 2024

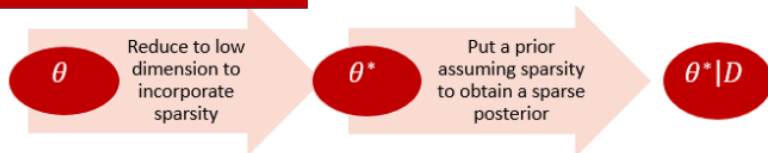
December 2024

NC STATE
UNIVERSITY

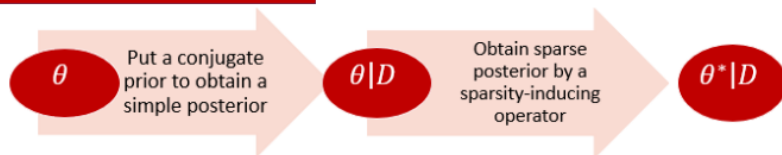
- **Problem:** $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ where $\mathbf{X} \in \mathbb{R}^n \times \mathbb{R}^p$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ for some $\sigma > 0$, $p > n$.
- Only a few predictors, say s_0 are active. The set S_0 comprises the active components.
- **Goals:**
 - ① **Optimal rates of posterior contraction** around the true coefficients utilizing the underlying low-dimensional structure.
 - ② **Accurate variable selection.**
 - ③ Construct credible intervals with **valid frequentist coverage**.

Motivation

Classical Bayesian Approach



New Bayesian Approach



- Analytical forms of posterior are typically not present when sparsity restrictions are incorporated in the prior.
- Coverage of the resulting credible regions is thus difficult to obtain.
- The procedure would be much easier to compute and study if the sparsity restrictions were not imposed before sampling from the posterior.
- Rather, a correction map could convert the posterior samples to the desired sparse space.

- **Conjugate Prior:**

$$\boldsymbol{\theta}|\sigma \sim \mathcal{N}_p\left(0, (\sigma^2/a_n)\mathbf{I}_p\right).$$

- **Posterior Distribution:**

$$\boldsymbol{\theta} | (\mathbf{X}, \mathbf{Y}, \sigma) \sim \mathcal{N}_p\left((\mathbf{X}^T \mathbf{X} + a_n \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y}, \sigma^2 (\mathbf{X}^T \mathbf{X} + a_n \mathbf{I}_p)^{-1}\right).$$

- **Sparsity-inducing map:**

$$\boldsymbol{\theta} \mapsto \boldsymbol{\theta}^* : \boldsymbol{\theta}^* = \arg \min_{\boldsymbol{u}} \{n^{-1} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{u}\|^2 + \lambda_n \|\boldsymbol{u}\|_1\}.$$

- The mapping can be simultaneously done with different tuning parameters as per the requirement.
- Different immersion posteriors may be used for different inferential goals.

Method: Tackling the noise variance σ^2

- An inverse-Gamma prior is conjugate in this setup. However, the resulting posterior is inconsistent.
- This is because, the posterior mean diverges as $n \rightarrow \infty$, that is, $\mathbb{E}(\tau|\mathbf{Y}) = n/\mathbf{Y}^T(\mathbf{I} - \mathbf{H}(a_n))\mathbf{Y} \rightarrow \infty$.
- This is rectified using an immersion map ι given by

$$\iota : \mathcal{T} \mapsto \mathcal{T}^* = \kappa\mathcal{T}$$

- This results in a posterior mean $\mathbb{E}(\kappa\mathcal{T}|\mathbf{Y}) = n\kappa/\mathbf{Y}^T(\mathbf{I} - \mathbf{H}(a_n))\mathbf{Y}$.
- Now, we choose

$$\kappa = n^{-1}\mathbf{Y}^T(\mathbf{I} - \mathbf{H}(a_n))\mathbf{Y}/\tilde{\sigma}^2$$

where $\tilde{\sigma}^2$ is a consistent estimator of σ^2 .

- The posterior mean is now $1/\tilde{\sigma}^2$ and the posterior variance is $2/(n\tilde{\sigma}^4) \rightarrow 0$ as $n \rightarrow \infty$ since $\tilde{\sigma}^2$ converges to the finite positive constant σ_0^2 .

Algorithm 1 Projection-posterior for θ

Step 1: Draw σ from the marginal posterior for σ and let

$$\sigma^* = \iota(\sigma)$$

be the image under the immersion map for σ .

Step 2: Draw θ from

$$\mathcal{N}_p((\mathbf{X}^T \mathbf{X} + a_n \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y}, (\sigma^*)^2 (\mathbf{X}^T \mathbf{X} + a_n \mathbf{I}_p)^{-1})$$

Step 3: Map to θ^* through the sparse projection map

$$\theta^* = \arg \min_{\mathbf{u}} \{n^{-1} \|\mathbf{X}\theta - \mathbf{X}\mathbf{u}\|^2 + \lambda_n \|\mathbf{u}\|_1\}$$

Theoretical Results: Optimal Rate of Recovery and Prediction

Theorem (Estimation and Prediction Consistency Rates)

For $\lambda_n \asymp \sqrt{\frac{\log(p)}{n}}$, under the compatibility, bounded true mean and non-collinearity conditions, we have for every sequence $M_n \rightarrow \infty$,

$$\mathbb{P} \left(\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^0\|_1 \geq M_{1n} s_0 \lambda_n \middle| \mathbf{Y} \right) \rightarrow 0$$

$$\mathbb{P} \left(\frac{1}{n} \|\mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^0)\|_2^2 \geq M_n s_0 \lambda_n^2 \middle| \mathbf{Y} \right) \rightarrow 0$$

in probability.

Theorem (Model selection consistency)

Suppose there exists constants b_3, b_4 satisfying $0 \leq b_3 < b_4 < b_2 - b_1$ for which $\lambda_n \propto n^{\frac{1+b_4}{2}}$ and $p_n = \mathcal{O}(e^{n^{b_3}})$. Then, under the beta-min, non-singularity and irrepresentability conditions,

$$\Pi(\theta^*(\lambda_n) =_s \theta^0 | \mathbf{Y}) \rightarrow 1$$

in probability.

Theoretical Results: Uncertainty Quantification by Debiasing

The j th component of a debiased/desparsified projection, say θ^{**} , for $j = 1, \dots, p$ is then obtained as follows.

$$\theta_j^{**} = \theta_j^* + \frac{\mathbf{R}^{(j)\top}(\mathbf{X}\theta - \mathbf{X}\theta^*)}{\mathbf{R}^{(j)\top}\mathbf{X}^{(j)}} = \frac{(\mathbf{X}\theta)^\top \mathbf{R}^{(j)}}{\mathbf{X}^{(j)\top} \mathbf{R}^{(j)}} - \sum_{k \neq j} P_{jk} \theta_k^*$$

where $\hat{\gamma}^{(j)} = \arg \min_{\gamma \in \mathbb{R}^{(p-1)}} \|\mathbf{X}^{(j)} - \mathbf{X}^{(-j)}\gamma\|_2^2/n + \lambda_j^{\mathbf{X}} \|\gamma\|_1$ and $\mathbf{R}^{(j)} = \mathbf{X}^{(j)} - \mathbf{X}^{(-j)}\hat{\gamma}^{(j)}$.

Theoretical Results: Coverage of Component-wise Credible Region

Theorem (Component-wise coverage)

Assuming the non-collinearity, bounded true mean and compatibility conditions hold uniformly in j , $\lambda_j^{\mathbf{X}} \asymp \sqrt{(\log p)/n}$ uniformly in j , and sparsity is of the order $s_0 = o(\sqrt{n}/\log p)$,

$$\sup_B \left| \mathbb{P} \left(\sigma^{-1} \sqrt{n} (\theta_j^{**} - \theta_j^0) \in B \mid \mathbf{Y} \right) - \mathcal{N}_p(B; m_j, \Sigma_{jj}) \right| \rightarrow 0$$

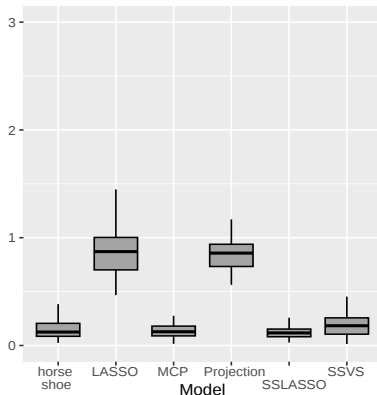
Define $C_j = \left\{ \theta_j^{**} : |\theta_j^{**} - m_j| < \Phi^{-1}(1 - \alpha/2) \frac{\sigma \sqrt{\mathbf{R}_{(j)}^T \mathbf{H}(a_n) \mathbf{R}_{(j)}}}{|\mathbf{X}_{(j)}^T \mathbf{R}_{(j)}|} \right\}$. Then,
 $\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_j^0} \left(\theta_j^0 \in C_j \right) = 1 - \alpha$ for all $j = 1, 2, \dots, p$.

Simulation Study: Setup

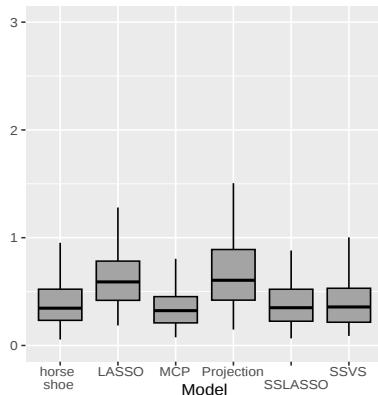
- We report the case where $p = 2000 > n = 100$.
- $s_0 = 10$.
- Here, two types of design are explored:
 - ① $\mathbf{X}$ consists of all independent covariates ($\Sigma = \sigma^2 \mathbf{I}$, $\sigma^2 = 1$ where $\mathbf{I}$ is the $p \times p$ identity matrix)
 - ② Autoregressive correlation structure ($\Sigma_{ij} = \rho^{|j-i|}$, where $\rho = 0.7$)
- Signal strength is taken to be $\beta_j = 2 \ \forall \ j \in S_0$

Simulation Study: Results

Boxplot of the MSE under independent design with $n = 100$, $p = 2000$



Boxplot of the MSE under correlated design with $n = 100$, $p = 2000$



Simulation Study: Results

(n, p, s_0)	Methods	Uncorrelated Design			Correlated Design		
		TPR	FDP	MCC	TPR	FDP	MCC
$n = 100$ $p = 2000$ $s_0 = 10$	LASSO	1	0.154 (0.183)	0.913 (0.114)	1	0.145 (0.017)	0.918 (0.105)
	MCP	1	0.056 (0.125)	0.958 (0.072)	1	0.056 (0.015)	0.969 (0.065)
	Horseshoe	1	0.013 (0.047)	0.990 (0.025)	1	0.015 (0.047)	0.992 (0.025)
	SSVS	1	0	1	1	0	1
	SSLASSO	1	0	1	0.995 (0.001)	0.005 (0.050)	0.995 (0.046)
	Projection	1	0.084 (0.063)	0.944 (0.108)	1	0.018 (0.037)	0.991 (0.019)

Table: Variable selection performances of the different methods when sample size $n = 100$ and dimension $p = 2000$. The mean (sd) of $M = 100$ replications are reported.

Simulation Study: Results

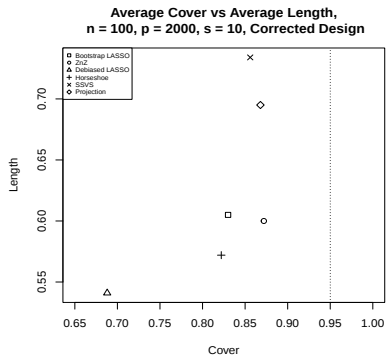
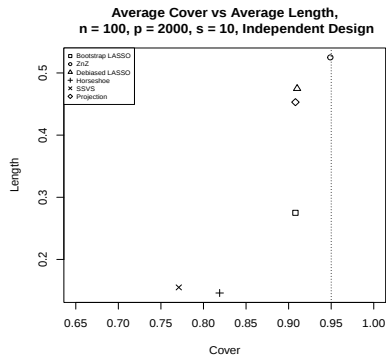
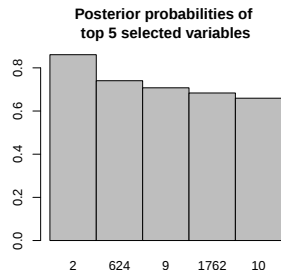
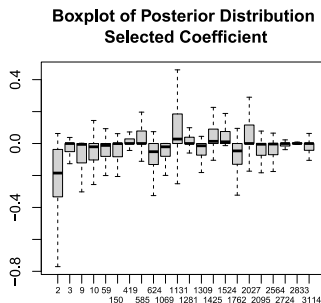


Figure: The average coverage probabilities for the 10 signals are plotted against the average lengths of the confidence/credible intervals for the competing methods.

Real Data Analysis: riboflavin data

- Response of interest: Logarithm of the riboflavin production rate.
- Predictors: Logarithm of the expression levels of 4088 genes.
- Sample size: 71 subjects.
- We studied the prediction errors in 10 random splits of the data, with 60 data points in the train set and the remaining 11 observations in the test set.
- The average prediction errors for LASSO for horseshoe were **97.43** and **85.33** respectively, and that for the sparse projection-posterior was **90.93**.

Real Data Analysis: riboflavin data



Top Selected Variables	2	624	9	1762	10
Debiased-projection credible interval	$(-0.32578, 0)$	$(-0.08465, 0)$	$(-0.26747, 0)$	$(-0.21748, 0)$	$(-0.22362, 0)$

Table: Credible Intervals using the debiased-projection posterior for the top 5 most selected variables by the sparse-projection posterior method.

The sparse projection-posterior approach...

- ...is one of the first Bayesian high-dimensional linear regression models to establish theoretical coverage guarantees,
- ...is a new take on Bayesian modeling, similar to the variation methods,
- ...is very flexible: choice of the loss and penalty functions in the optimization function, choice of the tuning parameter based on the inferential goal are all user-dependent,
- ...is compliant with distributed computing,
- ...and, is easily extendable to many other models!

Thank you!