Bayesian High-dimensional Linear Regression with Sparse Projection-posterior

Samhita Pal

Postdoctoral Research Scholar, Department of Biostatistics,
Vanderbilt University Medical Center

Joint work with Prof. Subhashis Ghosal

The 2026 ISBA World Meeting

Outline

Introduction

Method and Motivation

Some Technicalities

Theoretical Properties

Simulation Study

Introduction

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

Available Solutions

- **Regularization** adds a penalty to the objective function to encourage sparse solutions. Examples include:
  1. LASSO
  2. MCP
  3. SCAD
  4. Adaptive LASSO

Available Solutions

- **Regularization** adds a penalty to the objective function to encourage sparse solutions. Examples include:
  1. LASSO
  2. MCP
  3. SCAD
  4. Adaptive LASSO
- **Spike-and-slab priors** combine a point mass at zero with a diffuse slab for variable selection. SSVS provides a computationally faster posterior sampling approach.

Available Solutions

- **Regularization** adds a penalty to the objective function to encourage sparse solutions. Examples include:
  1. LASSO
  2. MCP
  3. SCAD
  4. Adaptive LASSO
- **Spike-and-slab priors** combine a point mass at zero with a diffuse slab for variable selection. SSVS provides a computationally faster posterior sampling approach.
- **Continuous shrinkage priors** use a single density with high concentration near zero and heavy tails. Examples include:
  1. Horseshoe prior
  2. Normal-gamma prior
  3. Double Pareto prior
  4. Dirichlet-Laplace prior

Proposed Solution: Immersion posterior

- The key idea is to first ignore the complication due to the structural restriction (here, sparsity) and use an 'unrestricted prior' Π on Θ so that the posterior updating $\Pi \mapsto \Pi(\cdot|\mathrm{Data})$ is simple.

Proposed Solution: Immersion posterior

- The key idea is to first ignore the complication due to the structural restriction (here, sparsity) and use an 'unrestricted prior' Π on Θ so that the posterior updating $\Pi \mapsto \Pi(\cdot|\mathrm{Data})$ is simple.
- Draw samples from the 'unrestricted posterior'.

Proposed Solution: Immersion posterior

- The key idea is to first ignore the complication due to the structural restriction (here, sparsity) and use an 'unrestricted prior' Π on Θ so that the posterior updating $\Pi \mapsto \Pi(\cdot|\mathrm{Data})$ is simple.
- Draw samples from the 'unrestricted posterior'.
- Then use a suitable "immersion map" $\iota : \Theta \to \Theta^*$ to convert unrestricted posterior samples to posterior samples complying with the structural restriction.

Proposed Solution: Immersion posterior

- The key idea is to first ignore the complication due to the structural restriction (here, sparsity) and use an 'unrestricted prior' Π on Θ so that the posterior updating $\Pi \mapsto \Pi(\cdot|\mathrm{Data})$ is simple.
- Draw samples from the 'unrestricted posterior'.
- Then use a suitable "immersion map" $\iota : \Theta \to \Theta^*$ to convert unrestricted posterior samples to posterior samples complying with the structural restriction.
- The resulting induced distribution Π^* is called the Immersion Posterior.

Outline

Introduction

Method and Motivation

Some Technicalities

Theoretical Properties

Simulation Study

Sparse Projection Posterior

- **Conjugate Prior**:
  $\boldsymbol{\theta}|\sigma \sim \mathscr{N}_p\left(\mathbf{0}, (\sigma^2/a_n)\boldsymbol{I}_p\right)$
- **Posterior Distribution**: $\boldsymbol{\theta}|(\mathbf{X}, \mathbf{Y}, \sigma) \sim$
  $\mathscr{N}_p\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + a_n\boldsymbol{I}_p\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + a_n\boldsymbol{I}_p\right)^{-1}\right)$.
- **Sparsity-inducing map**:
  $\boldsymbol{\theta} \mapsto \boldsymbol{\theta}^* : \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{u}}\{n^{-1}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{u}\|^2 + \lambda_n\|\boldsymbol{u}\|_1\}$.
- This induced posterior distribution of $\theta^{*(s)}$ is the **sparse projection posterior**.
- The mapping can be simultaneously done with different tuning parameters as per the requirement.
- Different projection posteriors may be used for different inferential goals.

Variational Bayes vs Sparse Projection Posterior

- **Variational Bayes (VB):** Restricts attention to a tractable family of distributions $q(\theta)$ and chooses q by minimizing a KL divergence to the full posterior $\Pi(\theta \mid Y)$. The optimization is at the *distribution level*.

Variational Bayes vs Sparse Projection Posterior

- **Variational Bayes (VB):** Restricts attention to a tractable family of distributions $q(\theta)$ and chooses q by minimizing a KL divergence to the full posterior $\Pi(\theta \mid Y)$. The optimization is at the *distribution level*.
- In VB, all inference (point estimates, credible sets) is based on $q(\theta)$, which approximates the entire posterior distribution in a simpler parametric form.

Variational Bayes vs Sparse Projection Posterior

- **Variational Bayes (VB):** Restricts attention to a tractable family of distributions $q(\theta)$ and chooses q by minimizing a KL divergence to the full posterior $\Pi(\theta \mid Y)$. The optimization is at the *distribution level*.
- In VB, all inference (point estimates, credible sets) is based on $q(\theta)$, which approximates the entire posterior distribution in a simpler parametric form.
- **Projection posterior:** Starts from the exact posterior samples $\{\theta^{(s)}\}$ from $\Pi(\theta \mid Y)$ and maps each draw to a constrained / penalized version $\theta^{*(s)} = \iota(\theta^{(s)})$.

Variational Bayes vs Sparse Projection Posterior

- **Variational Bayes (VB):** Restricts attention to a tractable family of distributions $q(\theta)$ and chooses q by minimizing a KL divergence to the full posterior $\Pi(\theta \mid Y)$. The optimization is at the *distribution level*.
- In VB, all inference (point estimates, credible sets) is based on $q(\theta)$, which approximates the entire posterior distribution in a simpler parametric form.
- **Projection posterior:** Starts from the exact posterior samples $\{\theta^{(s)}\}$ from $\Pi(\theta \mid Y)$ and maps each draw to a constrained / penalized version $\theta^{*(s)} = \iota(\theta^{(s)})$.
- Here the optimization is at the *parameter level* (per draw), often in a simple, non-iterative form.

Outline

Introduction

Method and Motivation

Some Technicalities

Theoretical Properties

Simulation Study

Method: Tackling the Noise Variance σ^2

- An inverse-Gamma prior is conjugate for the noise variance σ^2. However, the resulting posterior is inconsistent when $p > n$.
- Define $\boldsymbol{H}(a_n) := \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + a_n \boldsymbol{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}$
- Suppose we put the non-informative prior on the precision $\tau = \sigma^{-2}$, given by $p(\tau) \propto 1/\tau$.
- The posterior is then obtained as

$$\tau|\mathbf{Y} \sim \mathrm{Gamma}(n/2, \mathbf{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}(a_n))\mathbf{Y}/2).$$

- Consequently, the posterior mean diverges as $n \to \infty$, that is,

$$\mathbb{E}(\tau|\mathbf{Y}) = n/\mathbf{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}(a_n))\mathbf{Y} \to \infty.$$

Remedy

- This is rectified using an immersion map ι given by $\iota : \tau \mapsto \tau^* = \kappa\tau$
- This results in a Gamma posterior is

$$(\kappa\tau|\mathbf{Y}) \sim \text{Gamma}(n/2, \kappa^{-1}\mathbf{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}(a_n))\mathbf{Y}/2),$$

  with posterior mean $\mathbb{E}(\kappa\tau|\mathbf{Y}) = n\kappa/\mathbf{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}(a_n))\mathbf{Y}$.
- Now, we choose $\kappa = n^{-1}\mathbf{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}(a_n))\mathbf{Y}/\tilde{\sigma}^2$, 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

$$2/(n\tilde{\sigma}^4) \to 0 \text{ as } n \to \infty$$

  since $\tilde{\sigma}^2$ converges to the finite positive constant σ_0^2.
- Consequently, we can use the map

$$\sigma \mapsto \sigma^* = \frac{n^{1/2}\tilde{\sigma}}{(\mathbf{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}(a_n))\mathbf{Y})^{1/2}}\sigma.$$

The method at a glance

Algorithm 1 Projection-posterior for $\boldsymbol{\theta}$

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

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

be the image under the immersion map for σ.

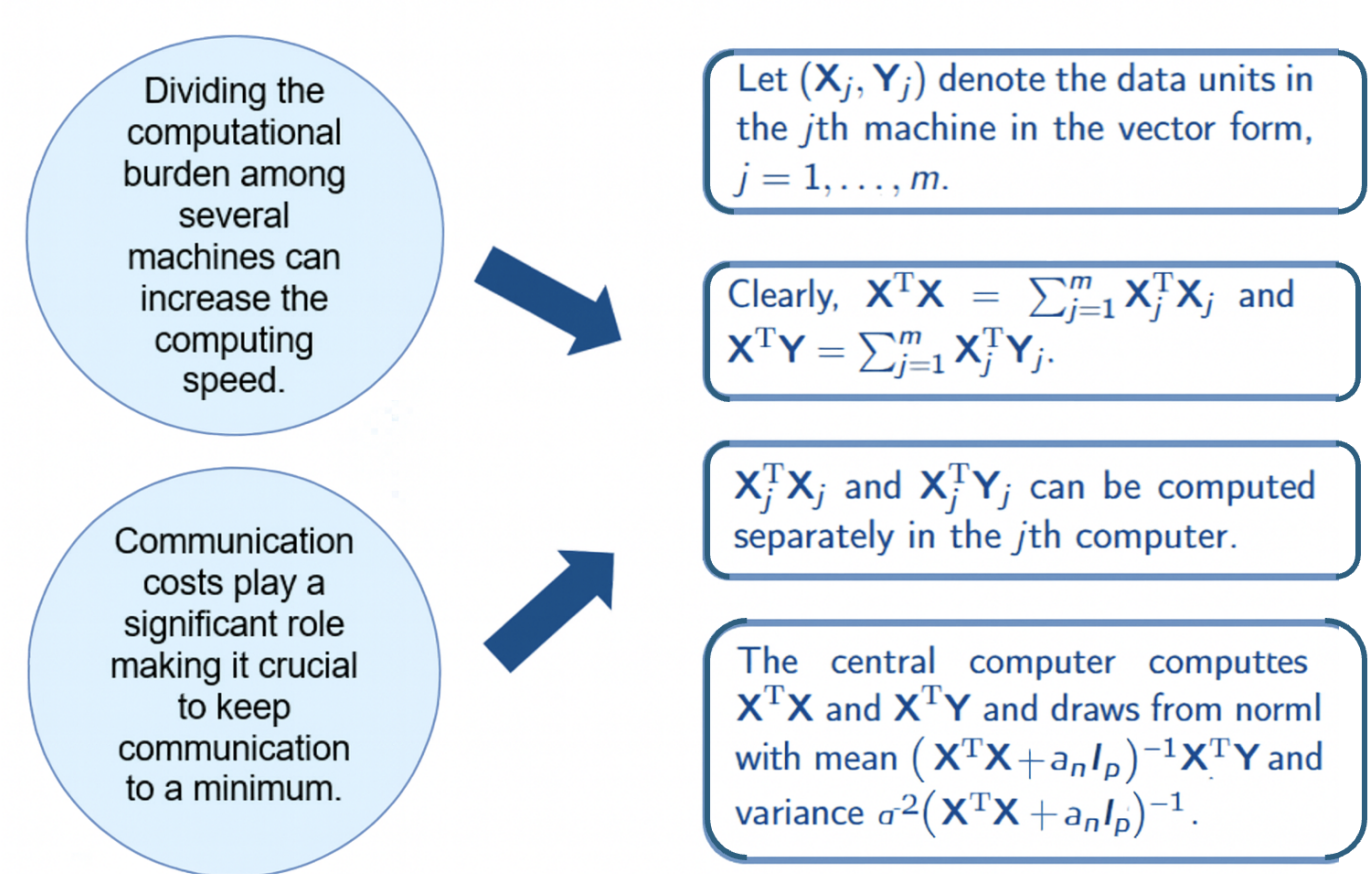
Step 2: Draw $\boldsymbol{\theta}$ from

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

Step 3: Map to $\boldsymbol{\theta}^*$ through the sparse projection map

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

Distributed Computing



Outline

Introduction

Method and Motivation

Some Technicalities

Theoretical Properties

Simulation Study

Assumptions

- **Compatibility condition**: The smallest eigenvalue in the restricted space $\{\boldsymbol{v} : \|\boldsymbol{v}_{S^c}\|_1 \leq L\|\boldsymbol{v}_S\|_1\}$ is lower bounded by a positive compatibility constant ϕ_0^2.
- **Beta-min condition**: The signal strengths of all the variables in the active set are larger than a specified value.
- **Non-singularity condition**: The eigenvalues of the design matrix corresponding to the relevant covariates are bounded from below.

Assumptions

- **Irrepresentability condition**: The total amount of dependence of the noise variables on the signals cannot be larger than the irrepresentability constant (or 1 in the weak case).
- **Non-collinearity condition**: Given the singular values $d_1, \ldots, d_n$ of $\mathbf{X}$, with $\text{rank}(\mathbf{X}) = n \ll p$, we assume that

$$\min_j d_j^2 \gg n a_n$$

- **Growth of the true signal**: Let $\|\mathbb{E}_{\boldsymbol{\theta}^0}(\mathbf{Y})\|_2 = \|\mathbf{X}\boldsymbol{\theta}^0\|_2 \leq b_n$ for some $b_n = o(d_{\min}^2 \sqrt{\log p}/a_n)$, where $d_{\min}^2 := \min\{d_k^2 : 1 \leq k \leq n\}$, and $d_1, \ldots, d_n$ are the singular values of $\mathbf{X}$.

Optimal Rate of Recovery and Prediction

Theorem (**Estimation and Prediction Consistency Rates**)

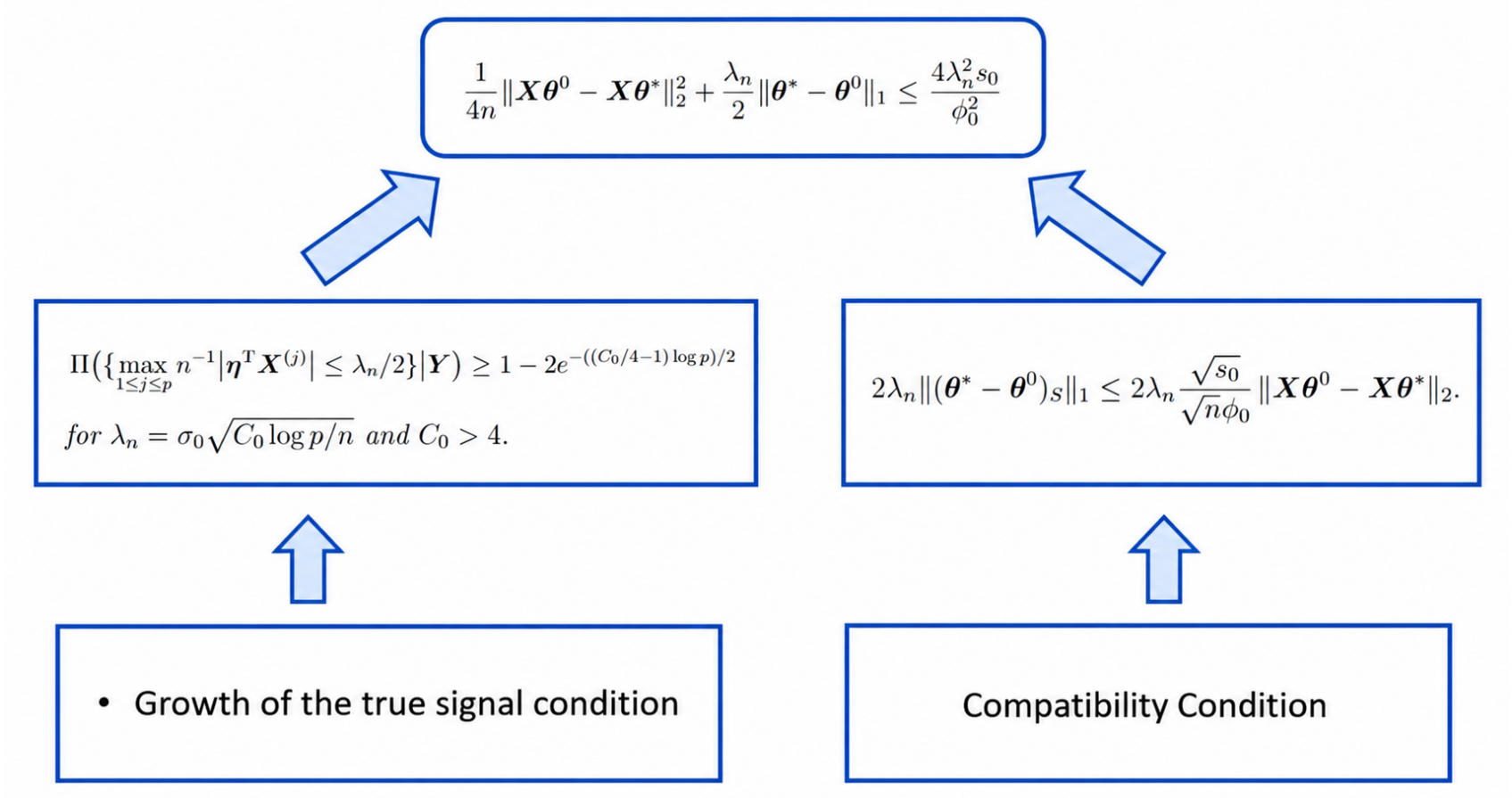
For $\lambda_n \asymp \sqrt{\frac{\log(p)}{n}}$, *under the compatibility, bounded true mean and non-collinearity conditions, we have for every sequence* $M_n \to \infty$,

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

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

in probability.

Proof Outline



Model Selection

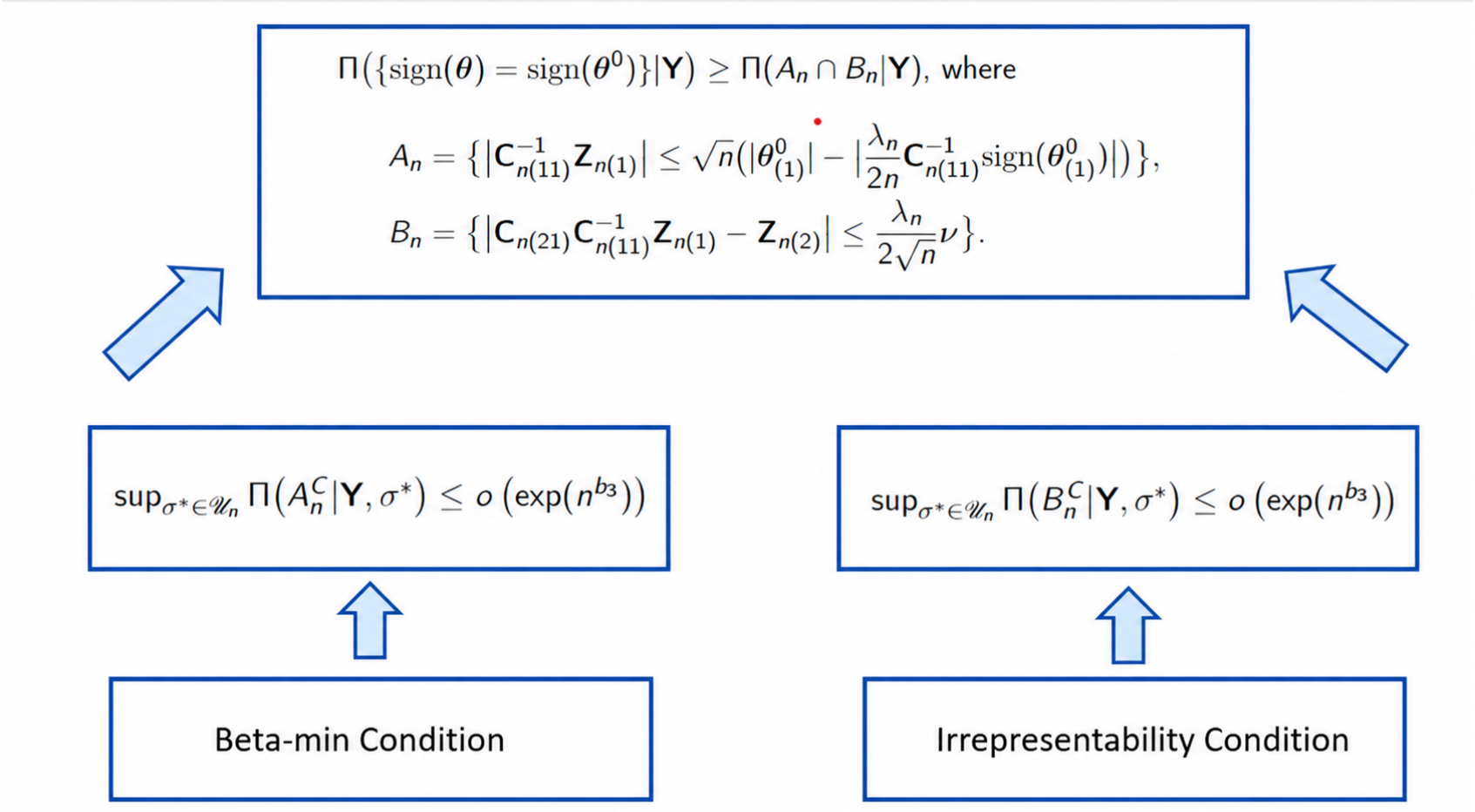
Theorem (**Model selection consistency**)

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* $p_n = \mathscr{O}(e^{n^{b_3}})$. *Then, under the beta-min, non-singularity and irrepresentability conditions,*

$$\Pi\left(\boldsymbol{\theta}^*(\lambda_n) =_s \boldsymbol{\theta}^0 | \mathbf{Y}\right) \to 1$$

in probability.

Proof Outline



Uncertainty Quantification by Debiasing

- The jth component of a debiased/desparsified projection, say $\boldsymbol{\theta}^{**}$, for $j = 1, \ldots, p$ is then obtained as follows.

$$\theta_j^{**} = \theta_j^* + \frac{\mathbf{R}^{(j)^{\mathrm{T}}}(\mathbf{X}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\theta}^*)}{\mathbf{R}^{(j)^{\mathrm{T}}}\mathbf{X}^{(j)}}$$

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

- Define $C_j = [\tilde{\theta}_j - q_{j,\alpha}, \tilde{\theta}_j + q_{j,\alpha}]$ to be the symmetric $(1-\alpha)$-credible interval for θ_j, $j = 1, \ldots, p$.
- $\tilde{\theta}_j$ is the median of the posterior distribution of θ_j^{**} and $q_{j,\alpha}$ is the $(1-\alpha/2)$-quantile of the posterior distribution of $|\theta_j^{**} - \tilde{\theta}_j|$.

Coverage of Component-wise Credible Region

Theorem (**Component-wise coverage**)

Assuming the non-collinearity, growth of the 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 $s_0 = o(\sqrt{n}/\log p)$,

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

Then, $\lim_{n \to \infty} \mathbb{P}_{\theta_j^0} \left(\theta_j^0 \in C_j \right) = 1 - \alpha$ *for all* $j = 1, 2, \ldots, p$.

Outline

Introduction

Method and Motivation

Some Technicalities

Theoretical Properties

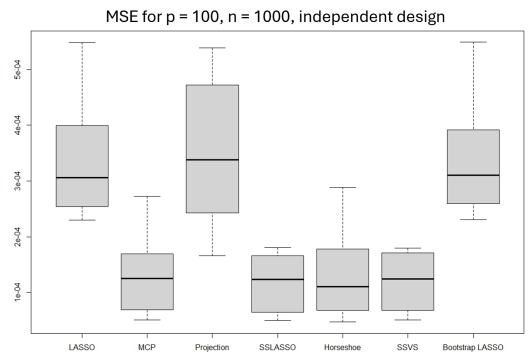
Simulation Study

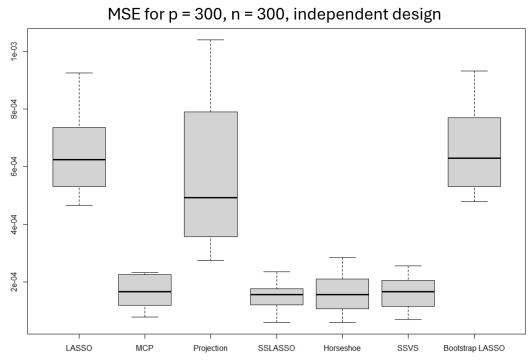
Simulation Setup

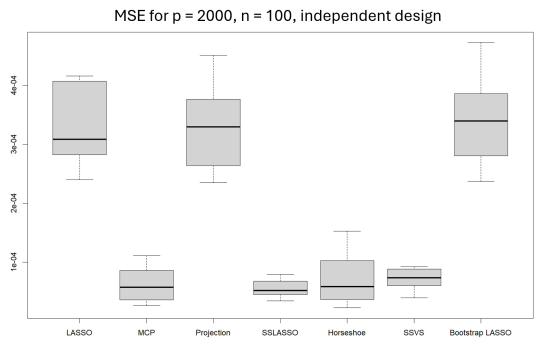
- $p = 100 < n = 1000$, $p = 300 = n$ and $p = 2000 > n = 100$.
- $s_0 = 10$.
- Independent and AR(1) with $\rho = 0.7$ covariates
- Signal strength: $\beta_j = 2 \ \forall \ j \in S_0$
- Projection-posterior method is compared with Bayesian LASSO, SSLASSO, horseshoe prior, MCP, LASSO, debiased LASSO, and bootstrap LASSO.
- For Bayesian methods 10000 MCMC samples were drawn after 2000 burn in.
- Projection method uses the tuning of the LASSO chosen by CV.

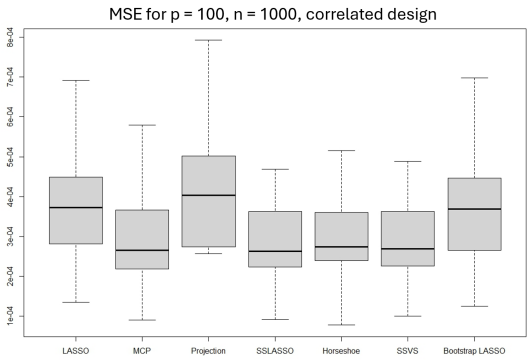
Estimation Performance

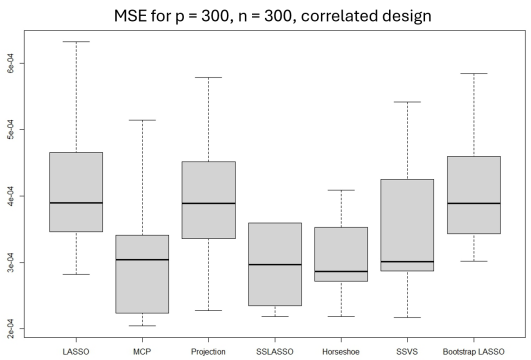
Mean Squared Estimation Error

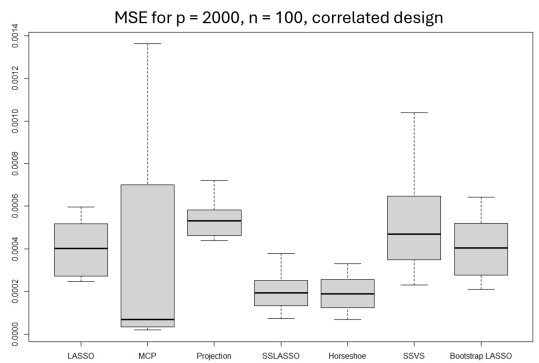




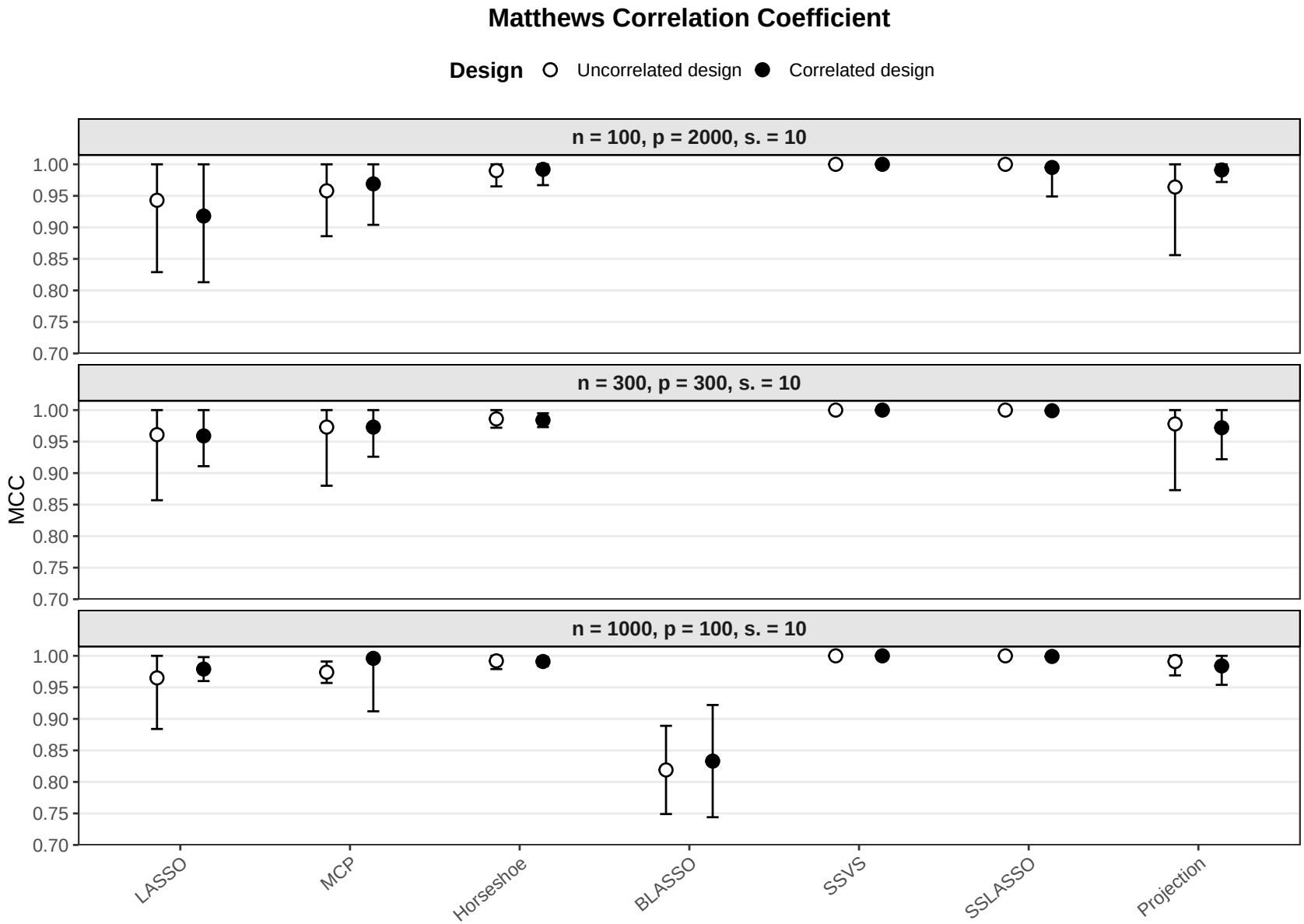




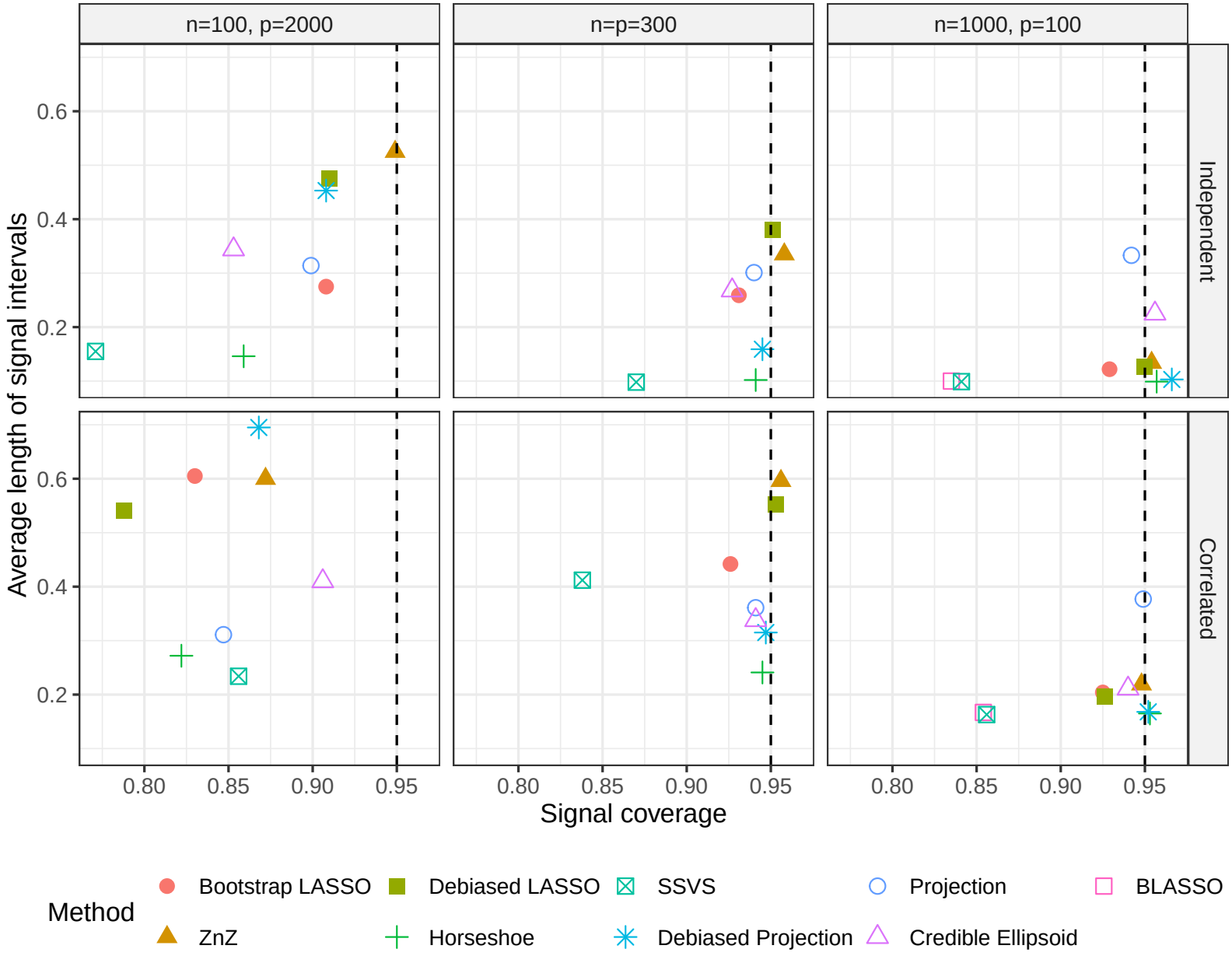




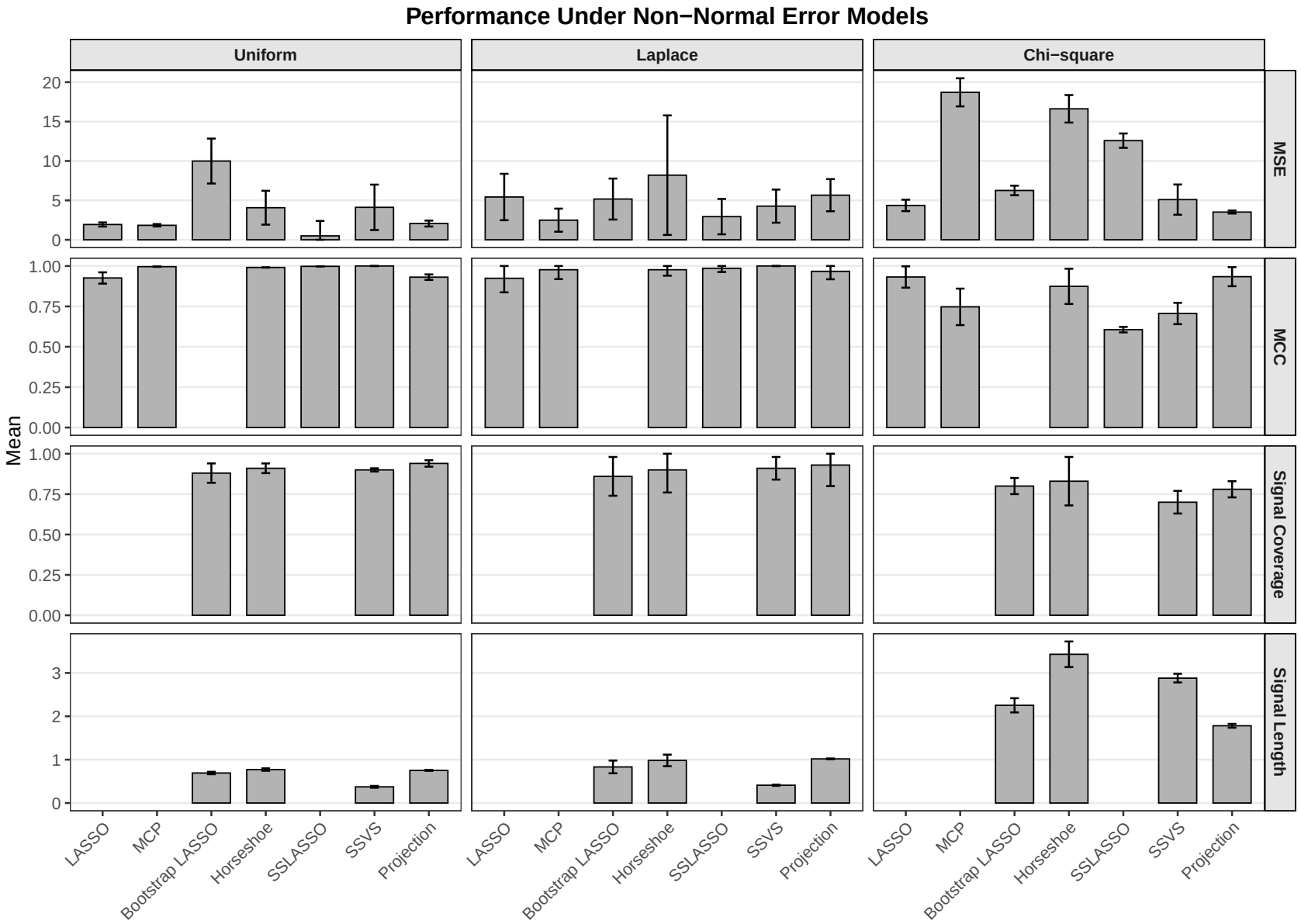
Model Selection Performance



Coverage Performance



Non-normal Error



Conclusion

- We introduced the **Sparse Projection Posterior**, which separates computationally convenient Bayesian posterior sampling from the enforcement of sparsity.
- The method provides:
  - optimal posterior contraction and prediction rates,
  - consistent variable selection, and
  - credible regions with asymptotically valid frequentist coverage.
- Posterior computation is simple and scalable, relying on conjugate posterior sampling followed by a standard penalized projection.
- The immersion-posterior framework also extends naturally to **grouped linear regression** through Group LASSO, Adaptive Group LASSO, and Group SCAD projection maps.

Thank You!

Joint credible ellipsoid

Model selection step

Let $S_0 \subset \{1, \ldots, p\}$ be the true active set of predictors with $|S_0| = s_0 \ll n$. Assume the conditions for selection consistency hold so that

$$\Pi(S = S_0 \mid \mathbf{Y}) \longrightarrow 1 \qquad \text{in probability.}$$

Let $\mathbf{X}_{(1)}$ denote the $n \times s_0$ submatrix of $\mathbf{X}$ containing the columns indexed by S_0.

Reduced model and posterior

On the selected model we work with the reduced regression

$$\mathbf{Y} = \mathbf{X}_{(1)} \boldsymbol{\theta}_{S_0} + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Let $(\boldsymbol{\theta}, \sigma)$ be drawn from the conjugate normal–inverse–gamma posterior under this reduced model, and denote the posterior mean on S_0 by $\hat{\boldsymbol{\theta}}^{\mathrm{R}}_{S_0}$.

Joint credible ellipsoid

Define the $(1-\alpha)$ joint credible ellipsoid B_n as

$$\left\{\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_{S_0}^\top, \mathbf{0}_{S_0^c}^\top) : (\boldsymbol{\theta}_{S_0} - \hat{\boldsymbol{\theta}}_{S_0}^{\mathrm{R}})^\top (\mathbf{X}_{(1)}^\top \mathbf{X}_{(1)} + a_n \mathbf{I}_{s_0})(\boldsymbol{\theta}_{S_0} - \hat{\boldsymbol{\theta}}_{S_0}^{\mathrm{R}}) \leq r_\alpha \right\},$$

where r_α is the $(1-\alpha)$ posterior quantile of the quadratic form

$$(\boldsymbol{\theta}_{S_0} - \hat{\boldsymbol{\theta}}_{S_0}^{\mathrm{R}})^\top (\mathbf{X}_{(1)}^\top \mathbf{X}_{(1)} + a_n \mathbf{I}_{s_0})(\boldsymbol{\theta}_{S_0} - \hat{\boldsymbol{\theta}}_{S_0}^{\mathrm{R}}),$$

computed from posterior draws under the reduced model.