Adaptive Bayesian Prediction Inference

Veronika Ročková

Booth School of Business
University of Chicago

Bayesian Prediction Inference

Predict $\boldsymbol{Y} \sim N(\boldsymbol{\beta}_0, r \times I_n)$ from $\boldsymbol{X} \sim N(\boldsymbol{\beta}_0, I_n)$ where

(1) $\boldsymbol{\beta}_0 \in \mathbb{R}^n$ is an *unknown* sparse mean vector where $\|\boldsymbol{\beta}_0\|_0 \leq s_n$

(2) $r \in (0, \infty)$ is known.

The goal is obtaining an *entire predictive density* $\hat{p}(\boldsymbol{y} \mid \boldsymbol{x})$ that is **close** to $\pi(\boldsymbol{y} \mid \boldsymbol{\beta}_0)$ in terms of the Kullback-Leibler loss

$$L(\boldsymbol{\beta}, \hat{p}(\cdot \mid \boldsymbol{x})) = \int \pi(\boldsymbol{y} \mid \boldsymbol{\beta}) \log \frac{\pi(\boldsymbol{y} \mid \boldsymbol{\beta})}{\hat{p}(\boldsymbol{y} \mid \boldsymbol{x})} \mathrm{d}\,\boldsymbol{y}, \tag{1}$$

We assess the quality of $\hat{p}(\cdot \mid \boldsymbol{x})$ by its risk

$$\rho(\boldsymbol{\beta}, \hat{p}) = \int \pi(\boldsymbol{x} \mid \boldsymbol{\beta}) L(\boldsymbol{\beta}, \hat{p}(\cdot \mid \boldsymbol{x})) \mathrm{d}\,\boldsymbol{x}.$$

Given a prior $\pi(\cdot)$, the average risk $r(\pi, \hat{p}) = \int \rho(\boldsymbol{\beta}, \hat{p}) \pi(\boldsymbol{\beta}) \mathrm{d}\,\boldsymbol{\beta}$ is minimized by the **Bayes (posterior) predictive density (BPD)**

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \int \pi(\boldsymbol{y} \mid \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid \boldsymbol{x}) \mathrm{d}\,\boldsymbol{\beta}. \tag{2}$$

Why Bayes?

Why integrate if we can just plug in?

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \pi(\boldsymbol{y} \mid \widehat{\boldsymbol{\beta}}) \tag{3}$$

In non-sparse setups, $\pi(\boldsymbol{y} \mid \widehat{\boldsymbol{\beta}}_{MLE})$ is *uniformly dominated* by BPD under the uniform prior.

By Jensen's inequality, BPD dominates a random plug-in estimator (3) when $\widehat{\boldsymbol{\beta}}$ is a random draw from the prior.

For sparse setups, Mukherjee and Johnstone (2015) quantified the minimax risk

$$\inf_{\hat{p}} \sup_{\boldsymbol{\beta}_0 : \|\boldsymbol{\beta}_0\| \leq s_n} \rho(\boldsymbol{\beta}, \hat{p}) \sim \frac{1}{1+r} s_n \log(n/s_n)$$

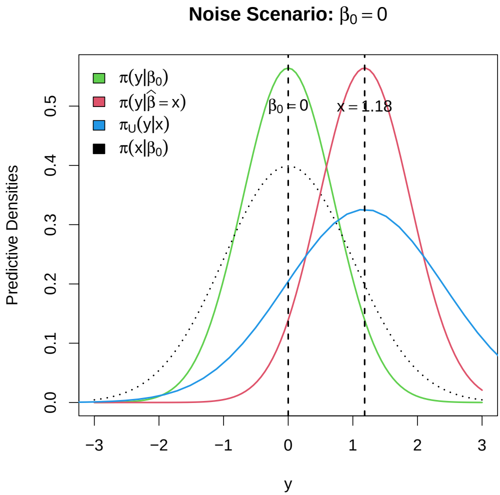
The minimax risk of plug-in density estimators (3) is

$$\frac{1}{r} \times s_n \log(n/s_n)$$

which is problematic for small r.

When $n = 1$...

Suppose $X \sim N(0, 1)$ and $r = 0.5$



RISK: Bayes: 1.07 and Plug-in: 1.306

REVIEW: Sparse Normal Means

Bayesian Estimation via posteriors under $\pi(\boldsymbol{\beta})$

Assume a product prior $\pi(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\beta_i)$

⇝ *Popular penalized-likelihood approach:* **LASSO**

$$\pi(\beta_i \mid \lambda) = Laplace(\beta_i \mid \lambda)$$

☹ Not as great properties
☺ Easy to compute (regression)

⇝ *Popular Bayesian approach:* **Spike-and-Slab**

$$\pi(\beta_i \mid \gamma_i) = \gamma_i \phi(\beta_i \mid \lambda_1) + (1 - \gamma_i)\delta_0(\beta_i), \quad \mathrm{P}(\gamma_i \mid \theta) = \theta, \quad \theta \sim \pi(\theta)$$

☺ Great properties: e.g. minimax rate $s_n \log(n/s_n)$ of posterior concentration
☹ Not so easy to compute (regression)

The Spike-and-Slab LASSO Prior (Rockova (2015))

A mixture of two LASSO priors with penalties λ_1 and λ_0

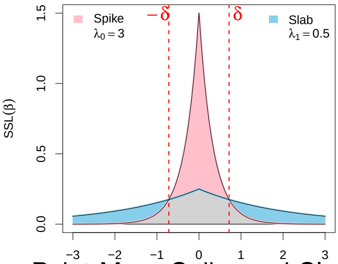
$$\pi_{SSL}(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} [\gamma_i \phi(\beta_i \mid \lambda_1) + (1 - \gamma_i)\phi(\beta_i \mid \lambda_0)]$$

$$\gamma_1, \ldots, \gamma_p \mid \theta \;\; iid \sim \texttt{Bern}(\theta), \quad \theta \sim \pi(\theta)$$

λ_1 **small**: slab distribution holds large coefficients steady

λ_0 **large**: spike distribution thresholds small coefficients

θ controls the sparsity



Converges to the Point-Mass Spike-and-Slab Prior as $\lambda_0 \to \infty$

Prediction Properties of Sparsity Priors

We inspect popular priors from a **predictive point of view**.

For independent product priors, BPD has a product form

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \hat{p}(y_i \mid x_i)$$

and

$$L(\boldsymbol{\beta}, \hat{p}(\cdot \mid \boldsymbol{x})) = \sum_{i=1}^{n} L(\beta_i, \hat{p}(\cdot \mid x_i)).$$

The predictive risk is additive and satisfies

$$(n - s_n)\rho(0, \hat{p}) < \rho(\boldsymbol{\beta}, \hat{p}) = \sum_{i=1}^{n} \rho(\beta_i, \hat{p}) \leq (n - s_n)\rho(0, \hat{p}) + s_n \sup_{\boldsymbol{\beta}_0 \in \mathbb{R}} \rho(\beta, \hat{p})$$

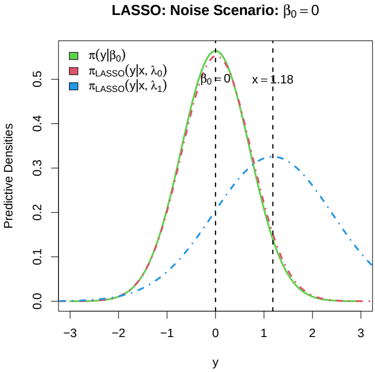
We need to control the risk at $\beta_0 = 0$ and, *at the same time*, when β_0 is large.

Bayesian LASSO

The Calibration Conflict

Two conflicting demands

$\lambda_0 = 10$ $\lambda_1 = 0.1$



λ_0 should be large for noise λ_1 should be small for signal

Bayesian LASSO Prediction Risk

We need:

$$\sup_{\beta\in\mathbb{R}} \rho(\beta,\hat{p}) \lesssim \frac{\log(n/s_n)}{1+r} \quad \textbf{and} \quad \rho(0,\hat{p}) \lesssim \frac{s_n \log(n/s_n)}{(n-s_n)(1+r)}$$

UPPER BOUND:

For $v = 1/(1+1/r)$ and Bayesian LASSO with $\lambda > 0$ we obtain

$$\rho(0,\hat{p}) \le \log\left(1 + \frac{\sqrt{2}}{\lambda\sqrt{\pi v}}\right) + \frac{4}{\lambda^2 v} \quad \text{and} \quad \sup_{\beta\neq 0} \rho(\beta,\hat{p}) \lesssim \lambda^2 + \frac{1}{\lambda^2}$$

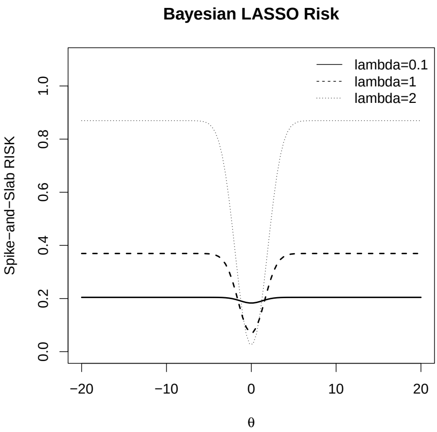
LOWER BOUND:

As $\lambda = \lambda_n \to \infty$ for some suitable $a > 0$

$$\inf_{\boldsymbol{\beta}\in\Theta(s_n)} \rho(\boldsymbol{\beta},\hat{p}) > (n-s_n)\left[(1-\Phi(a))\left(\frac{1}{\sqrt{v}} - 1\right)\frac{a}{2(\lambda_n + a)} - O(1/\lambda_n^2)\right].$$

↯ Traditional calibration $\lambda^2 \propto \log(n/s_n)$ does not work!

Bayesian LASSO Prediction Risk



Bayesian LASSO prediction risk $\rho(\beta, \hat{p})$ for $\lambda \in \{0.1, 1, 2\}$ and $r = 2$.

Spike-and-Slab LASSO
(fixed θ)

Spike-and-Slab Mixing of Predictive Densities

Consider a separable Spike-and-Slab prior for a **fixed** $\theta \in (0,1)$

$$\pi(\boldsymbol{\beta} \mid \lambda, \theta) = \prod_{i=1}^{n} \pi(\beta_i \mid \lambda, \theta), \quad \text{where } \pi(\beta \mid \lambda, \theta) = \theta \pi_1(\beta) + (1-\theta)\pi_0(\beta)$$

Denote by $m_j(x) = \int \pi(x \mid \mu)\pi_j(\mu)\mathrm{d}\,\mu$ the marginal likelihoods.

For $\theta \in (0,1)$, we define a mixing weight

$$\Delta_\theta(x) = \frac{\theta m_1(x)}{\theta m_1(x) + (1-\theta) m_0(x)} \tag{4}$$

BPD under the spike-and-slab prior is a mixture, i.e.

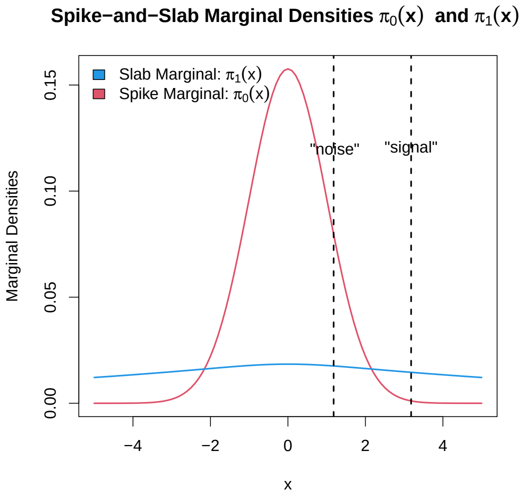
$$\hat{p}(y \mid x) = \Delta_\theta(x)\hat{p}_1(y \mid x) + [1 - \Delta_\theta(x)]\hat{p}_0(y \mid x) \tag{5}$$

where $\hat{p}_j(y \mid x) = \frac{\int \pi(y \mid \mu)\pi(x \mid \mu)\pi_j(\mu)\mathrm{d}\,\mu}{m_j(x)}$ for $j = 0, 1$ are BPD's under the spike/slab priors (respectively).

The Mixing Weight

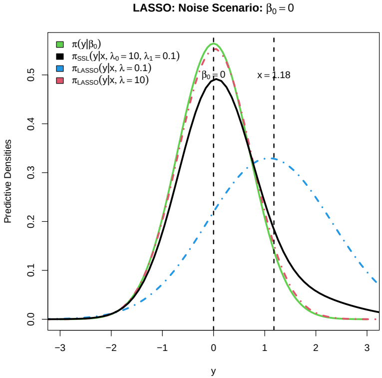
Denote by $BF(x;0,1)$ the Bayes factor for spike versus slab models

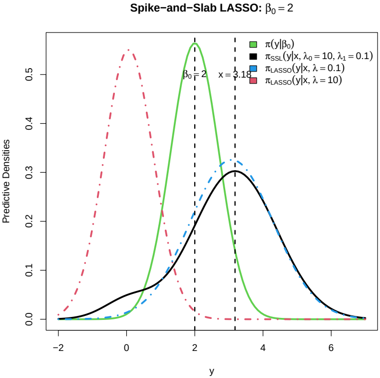
$$\Delta_\theta(x) = \frac{\theta m_1(x)}{\theta m_1(x) + (1-\theta)m_0(x)} = \left[1 + \frac{1-\theta}{\theta} BF(x;0,1)\right]^{-1}. \qquad (6)$$



Spike-and-Slab LASSO Predictive Densities

Adaptive mixing based on the magnitude of x.



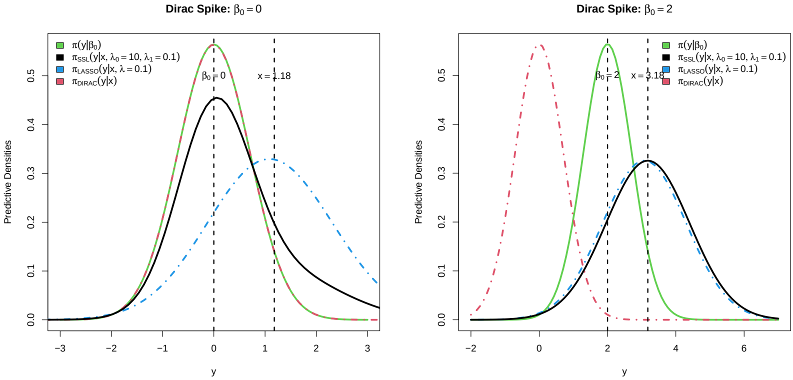


x is "small" and spike takes over over

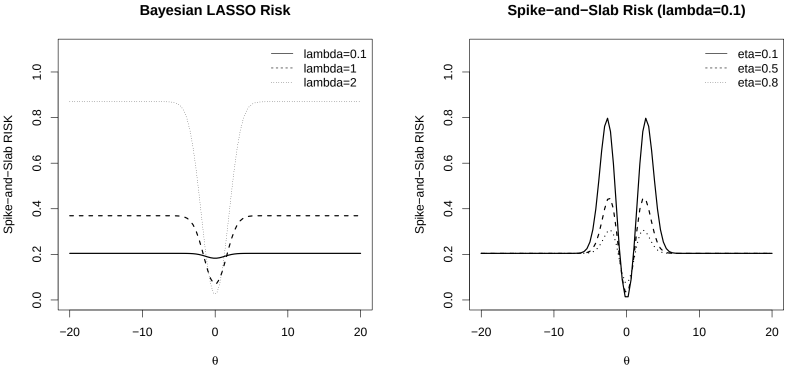
x is "large" and slab takes

Dirac Spike versus Laplace Spike

Laplace spike approximates Dirac spike $\pi_0(x) = \delta_0(x)$.



Dirac Spike and Laplace Slab



(Left) *Bayesian LASSO* prediction risk $\rho(\beta, \hat{p})$ for $\lambda \in \{0.1, 1, 2\}$;

(Right) *Spike-and-Slab* prediction risk for $\lambda = 0.1$ and $\theta \in \{0.1, 0.5, 0.8\}$;

Both plots correspond to $r = 2$.

Spike-and-Slab Priors are Rate-Minimax

Dirac Spike and Laplace Slab

Assume

$$(1-\theta)/\theta = n/s_n$$

and a Laplace slab where λ is fixed and depending on r.

With $s_n/n \to 0$ we have for any fixed $r \in (0, \infty)$

$$\sup_{\beta \in \Theta(s_n)} \rho(\beta, \hat{p}) \leq \frac{5}{1+r} s_n \log(n/s_n) + \widetilde{C}(r) \quad (7)$$

where $\widetilde{C}(r)$ a term depending on r.

- ☺ *Spike-and-Slab (Dirac version) is rate-minimax.*
- ☹ *Non-adaptive result! We need to know s_n to calibrate the prior!*

Spike-and-Slab Priors are Rate-Minimax

Spike-and-Slab LASSO: Laplace Spike and Laplace Slab

Assume

$$(1-\theta)/\theta = c$$

for some fixed constant $c > 0$.

Assume a Laplace spike with $\lambda_0 = n/s_n$ and Laplace slab with λ_1 fixed and depending on r.

With $s_n/n \to 0$ we have for any fixed $r \in (0,1)$

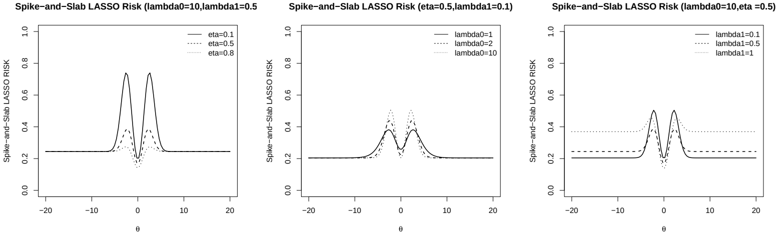
$$\sup_{\boldsymbol{\beta} \in \Theta(s_n)} \rho(\boldsymbol{\beta}, \hat{p}) \sim \frac{s_n}{1+r} \log(n/s_n). \tag{8}$$

The same conclusion holds for $r \in [1, \infty)$ for parameters $\boldsymbol{\theta} \in \Theta_n(s_n) \cap \{\boldsymbol{\theta} \in \mathbb{R}^n : \min_{1 \le i \le n} |\theta_i| > c_1 \sqrt{\log(n/s_n)}\}$ for suitable $c_1 > 0$.

☺ *Spike-and-Slab LASSO is rate-minimax.*

☹ *Non-adaptive result! We need to know s_n to calibrate the prior!*

Spike-and-Slab LASSO Prediction Risk



(Left) Varying θ for fixed $\lambda_0 = 10, \lambda_1 = 0.5$;
(Middle) Varying λ_0 for fixed $\theta = 0.1, \lambda_1 = 0.1$;
(Right) Varying λ_1 for fixed $\theta = 0.1, \lambda_0 = 10$.

Spike-and-Slab Priors
(random θ)

Random θ

Now we assume a *hierarchical* version (not an independent product)

$$\pi(\boldsymbol{\beta}) = \int_\theta \prod_{i=1}^{n} [(1-\theta)\delta_0 + \theta\pi_1(\beta_i)]\pi(\theta)\,\mathrm{d}\,\theta \quad \text{and} \quad \pi(\theta) \sim Beta(a,b) \quad (9)$$

for some $a, b > 0$.

We have

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \int_\theta \prod_{i=1}^{n} [\Delta_\theta(x_i)\hat{p}_1(y_i \mid x_i) + (1 - \Delta_\theta(x_i))\hat{p}_0(y_i \mid x_i)]\,\mathrm{d}\,\pi(\theta \mid \boldsymbol{x}),$$

and

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = E_{\theta \mid \boldsymbol{x}} \hat{p}(\boldsymbol{y} \mid \boldsymbol{x}, \theta). \quad (10)$$

The Kullback-Leibler loss of the predictive distribution under the hierarchical prior (9) satisfies

$$L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x})) \leq E_{\theta \mid \boldsymbol{x}} L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x}, \theta)).$$

Adapting to Sparsity s_n

The prediction risk under the hierarchical prior (9) satisfies for $\lambda > 2$

$$\rho(\boldsymbol{\beta}, \hat{p}) \leq s_n \left\{ C(\lambda, v) + (1 - v) \left[E_{\boldsymbol{x} \mid \boldsymbol{\beta}} E \log\left(\frac{1 - \theta}{\theta} \right) \mid \boldsymbol{x} \right] \right\}$$
$$+ D(n - s_n) \sup_{i:\beta_i \neq 0} E_{\boldsymbol{x}_{\setminus i} \mid \boldsymbol{\beta}} E \left(\frac{\theta}{1 - \theta} \mid \boldsymbol{x}_{\setminus i} \right).$$

for a suitable constant $C(\lambda, v) > 0$ and $D = 1 + 2/(a - 1)$, where $\boldsymbol{x}_{\setminus i}$ denotes the vector $\boldsymbol{x}$ without the i^{th} coordinate.

Adaptive minimax rate achieved when

$$E_{\boldsymbol{x} \mid \boldsymbol{\beta}} E \log\left(\frac{1 - \theta}{\theta} \right) \mid \boldsymbol{x} \lesssim \log(n/s_n)$$

and

$$\sup_{i:\beta_i \neq 0} E_{\boldsymbol{x}_{\setminus i} \mid \boldsymbol{\beta}} E \left(\frac{\theta}{1 - \theta} \mid \boldsymbol{x}_{\setminus i} \right) \lesssim s_n/n.$$

The Magic of Hierarchical Priors

Assume the hierarchical Spike-and-Slab prior (9) with $a, b > 0$.

Under the Gaussian model $\boldsymbol{X} \sim N_n(\beta, I)$, the posterior distribution $\pi(\beta \mid \boldsymbol{x})$ satisfies for any $\beta \in \Theta(s_n)$ with $s_n(\beta) = \|\beta\|_0$

$$E\left(\frac{\theta}{1-\theta} \mid \boldsymbol{x}\right) \leq \frac{a + E[s_n(\beta) \mid \boldsymbol{x}] + 1}{b-1}$$

and

$$E\left(\frac{1-\theta}{\theta} \mid \boldsymbol{x}\right) \leq E\left(\frac{b+n}{s_n(\beta) + a - 1} \mid \boldsymbol{x}\right).$$

Suggested calibration

$$a = 2 \qquad \text{and} \qquad b = n.$$

It is important that the posterior:

$E[s_n(\beta) \mid \boldsymbol{x}]$ does not overshoot s_n by too much.

$E[1/s_n(\beta) \mid \boldsymbol{x}]$ does not overshoot $1/s_n$ by too much.

The Posterior Does not Overshoot

Assume $\boldsymbol{X} \sim N_n(\beta_0, I)$ and the hierarchical Spike-and-Slab prior (9) with $a = 2$ and $b = n + 1$. Then for some suitable $M > 0$ we have

$$\sup_{\boldsymbol{\beta}_0 \in \Theta_n(s_n)} E_{\boldsymbol{x} \mid \boldsymbol{\beta}_0} E\left(\frac{\theta}{1-\theta} \mid \boldsymbol{x} \right) \leq M s_n / n + o(1) \quad \text{as } n \to \infty.$$

This result follows from Castillo and van der Vaart (2012).

This takes care of the noise coordinates in the risk upper bound:

$$D(n - s_n) \sup_{i:\beta_i \neq 0} E_{\boldsymbol{x}_{\setminus i} \mid \boldsymbol{\beta}} E\left(\frac{\theta}{1-\theta} \mid \boldsymbol{x}_{\setminus i} \right) \lesssim (n - s_n) \frac{s_n}{n}$$

The Posterior Does not Undershoot

Define

$$\Theta_n(s_n, \widetilde{M}) = \Theta_n(s_n) \cap \left\{ \boldsymbol{\beta} \in \mathbb{R}^n : \min_{i:\beta_i \neq 0} |\beta_i| > \widetilde{M}\sqrt{\log n} \right\}. \qquad (11)$$

Assume $\boldsymbol{X} \sim N_n(\boldsymbol{\beta}_0, I)$ and the hierarchical Spike-and-Slab prior (9) with $a = 2$ and $b = n + 1$.

Denote with S an index of all subsets of $\{1, \dots, n\}$ and define $c = (\widetilde{M}^2 - 2)/4$.

We have

$$\sup_{\boldsymbol{\beta} \in \beta_n(s_n, \widetilde{M})} P\Big(\exists j \text{ such that } \beta_j \neq 0 \quad \text{and} \quad j \notin S \,|\, \boldsymbol{x} \Big) \leq \frac{C e^{\lambda^2/2} s_n}{n^{c-1}}$$

with probability at least $1 - 2/n$. Assume $\lambda > 0$ such that $\lambda^2 \leq 2d \log n$ for some $d > 0$. Then for $c > 2 + d$ we have

$$\sup_{\boldsymbol{\beta} \in \Theta_n(s_n, \widetilde{M})} E_{\boldsymbol{x} \,|\, \boldsymbol{\beta}} E\left(\frac{1 - \theta}{\theta} \,|\, \boldsymbol{x} \right) \lesssim n/s_n.$$

... and finally!

Hierarchical Spike-and-Slab Prior

Assume the hierarchical prior (9) with a Laplace slab and with $a = 2$ and $b = n + 1$.

A bit of calibration needed for λ

Choose $\lambda^2 = vC_r$ for $C_r > 2/[v(1/2 + 4)]$ such that $\lambda > 2$ when $0 < r < 1$ and $\lambda^2 = (1 - v)C_r^*$ for $C_r^* > 2/[5(1 - v)]$ such that $\lambda > 2$ when $r \geq 1$.

Beta-min condition to get the minimax rate without a log factor

Denote $c = (\widetilde{M}^2 - 2)/4$ where $\widetilde{M}$ is the signal-strength constant in (11) then we have for $c > 2$

$$\sup_{\boldsymbol{\beta}_0 \in \Theta_n(s_n, \widetilde{M})} \rho(\boldsymbol{\beta}_0, \hat{p}) \lesssim \frac{s_n}{r+1} \log(n/s_n) \quad \text{and} \quad \sup_{\boldsymbol{\beta}_0 \in \Theta_n(s_n)} \rho(\boldsymbol{\beta}_0, \hat{p}) \lesssim \frac{s_n}{r+1} \log(n).$$

Hooray! Adaptive minimax rate (no knowledge of s_n required)!

Spike-and-Slab priors are great!

Rockova, V. *"Adaptive Bayesian Prediction Inference"* (Submitted (2023))

Rockova, V. *"Bayesian Estimation of Sparse Signals with Continuous Spike-and-Slab Priors"* (AoS (2018))

Rockova, V. and George (2016) *"The Spike-and-Slab LASSO"* (JASA (2016))