Lecture 8: Bayesian methods in high dimensions

Thibault Randrianarisoa

UTSC

March 26, 2026

Recap: Classical Posterior

- In conjugate models, can express the posterior in simple form (a multivariate Gaussian for instance)
- In more complex settings, can approximate posterior using some tractable class of distributions (variational Bayes) or sampling methods (MCMC)
- Large sample Gaussian approximations: the Bernstein-von Mises theorem guarantees that

$$\pi(\theta \mid \mathbf{X}^{(n)}) \approx \mathcal{N}(\hat{\mu}_n, \Sigma_n)$$

  as long as the sample size n is large enough, likelihood smooth & differentiable, true value θ_0 in interior of parameter space
- Importantly, it assumes that n is large relative to # parameters p (where $\theta \in \mathbb{R}^p$). What about high dimensional models where $p \gg n$?

Some High Dimensional Statistical Models

The model $\mathcal{P} = \{P_\theta^{(n)} : \theta \in \Theta\}$ can be

- Normal mean: $Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \sigma^2)$, $i = 1, \ldots, n$. Here $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ and $n = p$.
- Linear regression: $Y_i = \theta^T X_i + \varepsilon_i$, independent errors (possibly normal) with variance σ^2, $i = 1, \ldots, n$, $\theta \in \mathbb{R}^p$, possibly $p \gg n$.
- Normal covariance (or precision): $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$, $i = 1, \ldots, n$, with $\theta = \Sigma \in \mathbb{R}^{p \times p}$, possibly $p \gg n$.
- Exponential family: $X_i \overset{\text{iid}}{\sim} \text{ExpFamily}(\theta)$, $\theta \in \mathbb{R}^p$, possibly $p \gg n$.

Sparsity

In these situations, meaningful inference is possible only if there is a hidden lower-dimensional structure involving far fewer parameters. We call this **sparsity**.

- Normal mean: Only $s \ll n$ means θ_i are non-zero.
- Linear regression: Only $s \ll \min(p, n)$ coefficients θ_i are non-zero.
- Normal covariance (or precision):
  - (Nearly) banding structure: Total contribution of off-diagonal elements outside a band is small
  - Graphical model structure: Off-diagonal elements are non-zero only if the corresponding edges are connected.

Oracle

If sparsity structure is known, then inference reduces to the classical situation, and hence optimal procedures exist.

- For instance, in the normal mean model, if we knew which θ_i's are non-zero, we just estimate them.

The goal is to match the performance of the oracle within a small extra cost (which may come in the form of an additional log factor in the convergence rate).

In addition, If signals are sufficiently strong, one also likes to discover the true sparsity structure up to small error (for instance, one likes to conclude, with probability tending to one, the estimated sparsity agrees with the true sparsity).

Variable Selection

This motivates the study of **Variable selection:** the process of identifying the most relevant variables in a model (from a larger set of predictors for instance).

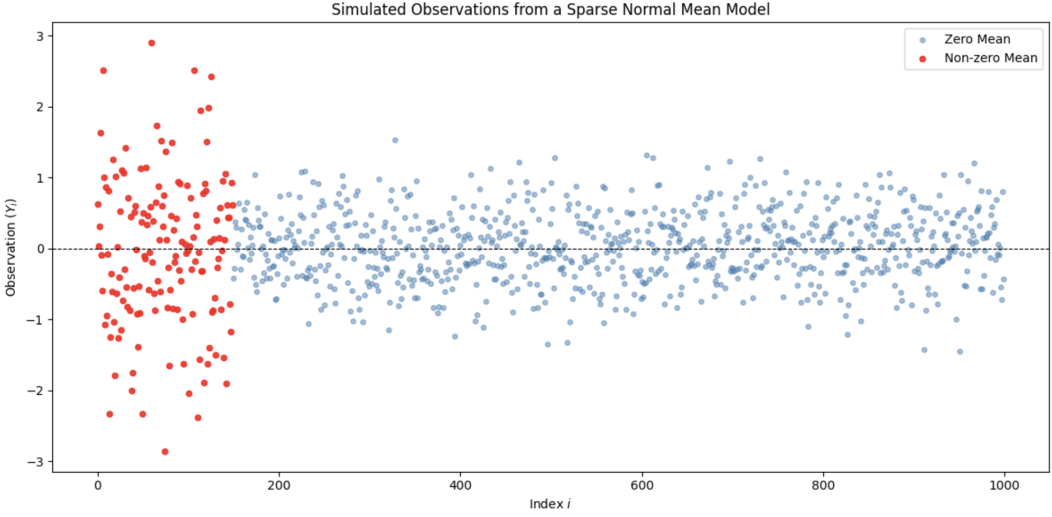
Assuming variables contribute unevenly to the outcome, we may want to identify the most "important" ones.

We can always include all available variables in a model, but reducing the number of variables helps:

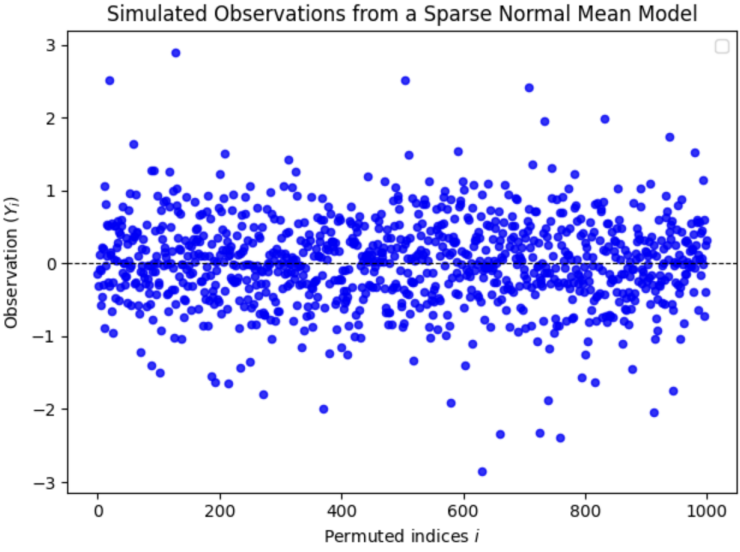
- reducing the computational/statistical complexity
- improving interpretability of the model

Also, evaluating all the potential subset of variables would have exponential complexity $O(2^p)$, which is intractable (i.e, empirical Bayes is not enough).

Normal means



Normal means



Penalization and LASSO

Numerous methods for estimating parameters in the high-dimensional setting have been proposed in the literature, most of which use penalization.

The idea is to add a suitable penalty term to a loss function to be optimized, ensuring the resulting solution is sparse. The most familiar method in linear regression is the LASSO for a linear regression model:

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \theta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\theta_j|.$$

If the noise has a Gaussian distribution $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, we already know it is equivalent to the MAP estimator under a Laplace prior.

We now focus on the Gaussian mean model (which is equivalent to linear regression with design matrix $X = I_n$).
Let $\|\beta\|_{2,n}^2 = n^{-1} \sum_{i=1}^n \beta_i^2$.

Convergence rate of Lasso

The LASSO estimator $\hat{\theta}$ with penalty $\lambda_n = O(\sqrt{\log(n)/n})$ satisfies

$$\|\hat{\theta} - \theta_0\|_{2,n} \leqslant C \left(\frac{s \log n}{n} \right)^{1/2}$$

with probability $1 - o(1)$.

The minimax rate of convergence for the above model is $\sqrt{s \log(n/s)/n}$, so the LASSO is almost optimal.

However, we are interested in Bayesian procedures and overall the performance of the whole prior (not just point estimates), via its contraction rate for instance.

In the sparse setup, the full posterior distribution $\Pi_\lambda^{\text{LASSO}}$ corresponding to the 'LASSO' prior with penalty factor λ does not contract at the same speed as its mode.

Theorem

For any $\lambda = \lambda_n$ such that $\sqrt{n}/\lambda_n \to \infty$, there exists $\delta > 0$ such that, as $n \to \infty$,

$$\Pi_{\lambda_n}^{\text{LASSO}}\left(\theta : \|\theta\|_{2,n} \leqslant \delta\sqrt{n}\left(\frac{1}{\lambda_n} \wedge 1\right) \,\middle|\, \mathbf{X}^{(n)}\right) \xrightarrow{\mathbb{P}} 0.$$

Intuition: The parameter λ in the Laplace prior faces conflicting demands: it must be large to shrink coefficients β_i to zero, but reasonable enough to model the nonzero coordinates.

Bayesian Inference in High-Dimensional Models

Good Bayesian procedures should possess desirable frequentist properties:

- Optimal rate of contraction
- Consistency in variable selection (or structure learning)
- Does a sparse version of the Bernstein-von Mises theorem hold, i.e., the posterior is asymptotically the product of normal of the oracle dimension and Dirac masses at zero?

Idea: A hidden lower-dimensional structure, such as sparsity, can be easily incorporated into a prior distribution (e.g., allowing a point mass at zero).

Spike-and-Slab Priors

Intuition: While an entry θ_i may likely be zero, the possibility of a non-zero or large value cannot be ruled out. We superimpose two regimes for a parameter θ_i: one for zero or minimal values, and one for possibly large values.

The spike-and-slab prior is formulated as a mixture distribution:

$$\pi(\theta) = \prod_{i=1}^{p} \pi(\theta_i)$$
$$\pi(\theta_i) = (1 - w)\phi_0(\theta_i) + w\phi_1(\theta_i)$$

with components:

- ϕ_0: the "spike", a density highly concentrated at 0 (like a point-mass δ_0 at 0).
- ϕ_1: the "slab", a density allowing intermediate and large values (usually symmetric about 0, like a Gaussian with large variance or Laplace density).
- w: a small mixing weight parameter that induces sparsity.

Recovery in ℓ_2 Norm

We consider a spike-and-slab prior with a Laplace slab with parameter λ and with a Beta$(1, n^u)$ hyperprior, $u > 1$, on the mixing weight w.

$$w \sim \text{Beta}(1, n^u)$$
$$\theta \sim \prod_{i=1}^{n} ((1-w)\delta_0 + w\text{Lap}(\lambda))$$

Theorem

If $\lambda = \sqrt{\log n}$, then for sufficiently large M,

$$\Pi\left(\theta : \|\theta - \theta_0\|_{2,n} > M\left(\frac{s \log n}{n}\right)^{1/2} \middle| \mathbf{X}^{(n)}\right) \xrightarrow{\mathbb{P}} 0.$$

Selection Consistency

The posterior distribution induces a distribution on the set $S_\theta \subset \{1, 2, \ldots, p\}$ on non-zero coordinates θ_i. It is desirable that this puts most of its mass on the true model S_0 corresponding to θ_0.

Theorem

For sufficiently large M,

$$\Pi\left(\theta : S_\theta \supset \left\{i : |\theta_{0i}| \geqslant M\sqrt{s \log n}\right\} \,\middle|\, \mathbf{X}^{(n)}\right) \xrightarrow{\mathbb{P}} 1,$$

and

$$\Pi\left(\theta : S_0 \subset S_\theta,\ S_0 \neq S_\theta \,\middle|\, \mathbf{X}^{(n)}\right) \xrightarrow{\mathbb{P}} 0,$$

As the support of a vector θ_0 is defined only in a qualitative manner by its coordinates θ_{0i} being zero or not, this will not be true in general.

Soft Selection: Continuous Shrinkage Priors

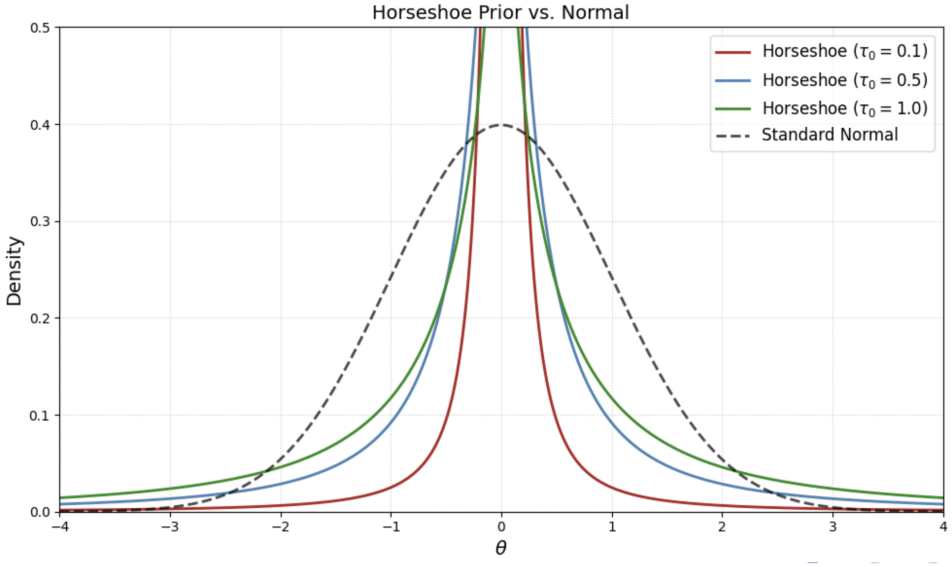
Issue: Hard spike-and-slab priors require exploring a massive combinatorial space (2^p possible models) via MCMC sampling, which often becomes computationally intractable in high dimensions.

A continuous alternative (to do a 'soft' selection): we can replace the discrete mixture with a single, continuous prior distribution that possesses a sharp peak at zero (mimicking the spike) and heavy tails (mimicking the slab).

The Horseshoe Prior:

$$\begin{aligned} \theta_i \mid \tau &\sim N(0, \tau^2) \\ \tau &\sim \text{Half-Cauchy}(0, \tau_0) \end{aligned}$$

Horseshoe Prior



We can also encode the sparsity-inducing behaviour in a variational family.

- **Mean-Field Variational Class:** We consider the variational class $\mathcal{Q}$:

$$\mathcal{Q} = \left\{ \bigotimes_{i=1}^{p} \{\gamma_i \mathcal{N}(\mu_i, \sigma_i^2) + (1-\gamma_i)\delta_0\},\ (\mu_i, \sigma_i^2, \gamma_i) \in \mathbb{R} \times \mathbb{R}^+ \times [0,1] \right\}$$

- **Theoretical Properties:** Based on the prior with a Laplace slabs, it:
  - Yields the same convergence rates and variable selection properties as exact spike-and-slab posteriors.
  - Laplace slabs are preferred in the prior, even though Gaussian slabs are used within $\mathcal{Q}$.

Computation:

- Solved via Coordinate Ascent Variational Inference (sequential updates of $\gamma_i, \mu_i, \sigma_i^2$).
- Main advantage: Significant computational speedup compared to standard MCMC.

⚠There are also some practical considerations:

- The CAVI output is sensitive to the order of parameter updates, requiring a prioritized update scheme.
- Requires careful tuning of the slab parameter λ.

Alternative construction

Issue: Encoding sparsity mechanisms directly in the prior makes posterior updating computationally and analytically complex.

Idea: Reverse the order. Perform posterior updating first using a simple, unconstrained prior (e.g., a conjugate normal prior), and then introduce sparsity.

Sparsity-Inducing Map: Sparsity is introduced at the posterior stage by applying a map to the unrestricted coefficients θ:

$$\theta \mapsto \theta^* := \arg\min \left\{ \|\theta - u\|^2 + \lambda_n P(u) : u \in \mathbb{R}^p \right\}$$

where $P(u)$ is a penalty function (like the ℓ_1-penalty $\|u\|_1$) and λ_n is a tuning parameter.

Inference via Projection-Posterior

The sparsity-inducing map is applied to *each sample* drawn from the easily-computed, unrestricted posterior.

The resulting distribution on θ^* is called the **sparse projection-posterior**.

Theoretical Properties:

- The projection-posterior concentrates near the true regression vector at the optimal rate.
- It is variable selection consistent under the same conditions required for the LASSO.

Immersion Posterior vs. Variational Bayes (VB)

Variational Bayes (VB):

- Optimization is performed on *distributions* (typically via iterative coordinate descent).
- The optimization procedure only needs to be done once.

Immersion Posterior:

- Optimization is in the *parameter space*, making the individual optimization problem considerably simpler (often non-iterative).
- However, the procedure must be performed for *each* individual sample drawn from the unrestricted posterior.