

Shrinkage Priors for Bayesian Inference – A Contextual Survey

Anandamayee Majumdar

Department of Mathematics
San Francisco State University, San Francisco, CA-94132, USA

Received: 02 July 2026; Revised: 14 September 2016; Accepted: 15 September 2026

Abstract

Shrinkage priors, alternatively called regularization priors, have become popular in modern Bayesian statistics, particularly in high-dimensional and sparse settings. Their evolution indicates a journey from decision-theoretic insights to structured hierarchical modeling. This survey provides a concise account of foundational contributions, emphasizing the problems each method addressed, the mathematical structure introduced, the practical trade-offs, limitations, and the intuition behind their behavior with practical and recent instances.

Key words: Bayesian model; Hierarchical model; Horseshoe; Lasso; Regularization; Ridge; Shrinkage.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Classical least squares becomes unstable or even inestimable once the number of predictors p approaches or exceeds the sample size n , and it offers no mechanism for exploiting the belief that only a handful of predictors actually matter. Shrinkage priors address both problems at once: by pulling coefficient estimates toward zero, they trade a small amount of bias for a substantial reduction in variance, which very often lowers overall estimation and prediction error relative to unregularized least squares. Bayesian shrinkage priors do not use fixed penalty, rather, a *prior* additionally yields a full posterior distribution rather than a single point estimate, giving coherent, automatic uncertainty quantification without a separate bootstrap or asymptotic approximation. Within the Bayesian framework, continuous shrinkage priors – and the global-local family in particular – are attractive because they combine this variance reduction and built-in inference with a shrinkage profile that adapts per coefficient: most coefficients are pulled sharply toward zero, while a minority of truly large effects can still be estimated with little attenuation, approximating the behavior of discrete variable selection without the combinatorial cost of searching over 2^p models.

Consider an example, the classical linear model:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I). \quad (1)$$

When p is large relative to n , the least squares estimator: $\hat{\beta} = (X^T X)^{-1} X^T y$ becomes unstable or undefined. Even when it exists, it can have extremely high variance. Regularization/shrinkage methods address this by introducing bias. Rather than estimating each parameter independently, regularization methods “borrow strength” across parameters.

This theme, first appeared formally in decision theory. Stein (Stein, 1956) showed that for $X \sim \mathcal{N}_p(\theta, I)$, the estimator $\hat{\theta} = X$ is inadmissible for $p \geq 3$. This was indeed shocking, since before this result unbiased estimators were widely considered optimal, no clear reason existed to introduce bias. Stein demonstrated that:

A biased estimator that shrinks toward zero can have strictly lower risk.

We show this with an example: Suppose $p = 10$ and $X_i \sim N(\theta_i, 1)$. The James–Stein estimator $\hat{\theta}_i^{JS} = \left(1 - \frac{8}{\|X\|^2}\right) X_i$ shrinks all coordinates toward zero. The Interpretation of this result implies that (i) if $\|X\|$ is small $\Rightarrow$ strong shrinkage, (ii) if $\|X\|$ is large $\Rightarrow$ weak shrinkage. This adaptive behavior is the conceptual ancestor of modern shrinkage priors.

With the advent of ridge regression, more shrinkage methods were eventually introduced and have been adopted in the Bayesian Statistical Literature. In this overview, the paper is organized as follows: Section 2, 3, 4, 5, discuss the different approaches used for shrinkage in some detail; Ridge regression, the Lasso, Scale Mixtures, and the Horseshoe. Section 6 discusses theoretical properties. Section 7 introduces simulation studies comparing some of these methods in terms of respective MSE and bias performances. We end the paper with a conclusion section.

2. Ridge regression

Ridge regression is a foundational method for stabilizing estimation in the presence of multicollinearity and high-dimensional predictors. Since its introduction by Hoerl and Kennard (Hoerl and Kennard, 1970), it has become a standard tool in statistical modeling, machine learning, and applied sciences.

The ridge estimator is defined as:

$$\hat{\beta} = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|_2^2 \right\}. \quad (2)$$

Ridge regression (Hoerl and Kennard, 1970) was introduced to address multicollinearity. This solved the problems of instability when predictors are highly correlated, as well as the exploding variance of least squares estimates.

The Bayesian interpretation of the Ridge regression corresponds to $\beta_j \sim \mathcal{N}(0, \tau^2)$.

Combining likelihood and prior yields a Gaussian posterior distribution:

$$\beta \mid y \sim \mathcal{N}(\mu_{\text{post}}, \Sigma_{\text{post}}), \quad (3)$$

with posterior mean given by:

$$\mu_{\text{post}} = (X^\top X + \lambda I)^{-1} X^\top y, \quad \lambda = \frac{\sigma^2}{\tau^2}. \quad (4)$$

This posterior mean coincides with the classical ridge estimator (Hoerl and Kennard, 1970).

Applications and limitations of ridge regression

This section highlights practical domains where ridge regression has been successfully applied, and where it has limitations.

In genomic prediction problems, the number of genetic markers often far exceeds the number of samples. Ridge regression is particularly well-suited to this setting because it assumes that many predictors contribute small effects rather than enforcing sparsity.

Meuwissen *et al.* (Meuwissen *et al.*, 2001) demonstrated that ridge-type models (often referred to as genomic BLUP) can accurately predict breeding values using dense marker maps. Similarly, de los Campos *et al.* (De Los Campos *et al.*, 2009) showed that ridge regression performs well in genome-wide prediction of quantitative traits. Ridge regression is widely used in genomic selection pipelines because it stabilizes estimation in $p \gg n$ regimes, captures polygenic effects, improves prediction accuracy for complex traits.

Financial datasets often contain highly correlated predictors such as asset returns, macroeconomic indicators, and factor models. Ordinary least squares performs poorly in this setting due to instability in covariance estimation.

Jagannathan and Ma (Jagannathan and Ma, 2003) showed that ridge-type regularization can improve portfolio allocation by reducing estimation error in the covariance matrix. DeMiguel *et al.* (DeMiguel *et al.*, 2009) further demonstrated that regularization/shrinkage-based approaches outperform naive diversification strategies in many empirical settings. Ridge regression is used to stabilize portfolio weights, reduce sensitivity to noisy covariance estimates, improve out-of-sample financial performance.

Ridge regression is a standard baseline in supervised learning due to its simplicity, convexity, and strong generalization properties. It is particularly effective when many features contribute weakly to the response.

Murphy (Murphy, 2012) describes ridge regression as a fundamental regularization method in probabilistic machine learning. It is also closely related to kernel ridge regression, which extends linear ridge regression to nonlinear feature spaces.

Saunders *et al.* (Saunders *et al.*, 1998) showed that ridge regression can be formulated efficiently in dual form, enabling scalable learning in high-dimensional feature spaces. Ridge regression is widely used because it provides strong predictive baselines, handles high-dimensional feature representations, extends naturally to nonlinear models via kernels.

In signal processing and applied mathematics, many estimation problems are ill-posed, meaning that small perturbations in data can lead to large changes in solutions. Ridge regression is equivalent to Tikhonov regularization, which stabilizes such inverse problems.

Tikhonov and Arsenin (Tikhonov and Arsenin, 1977) introduced this framework for solving ill-posed problems, and Hansen (Hansen, 1998) provided a comprehensive treatment of rank-deficient systems and regularization techniques. Ridge regression is used to stabilize ill-posed inverse problems, reconstruct signals from noisy observations, improve numerical conditioning.

Climate datasets often involve highly correlated variables such as temperature, pressure, and atmospheric composition. Ridge regression is useful in this context because it handles multicollinearity effectively. Mann *et al.* (Mann *et al.*, 1998) used regularized regression methods in reconstructing historical climate patterns from proxy data. Schneider (Schneider, 2001) further studied statistical methods for handling incomplete and noisy climate datasets using regularization/shrinkage approaches. Ridge regression helps stabilize long-term climate reconstructions, handle correlated environmental predictors, reduce overfitting in paleoclimate models.

Clinical datasets often contain correlated biomarkers, lab measurements, and demographic variables. Ridge regression is widely used for predictive modeling in medicine because it provides stable estimates and reduces overfitting. Van der Laan *et al.* (Van der Laan *et al.*, 2007) describe ensemble learning frameworks where ridge regression serves as a stable base learner. Steyerberg (Steyerberg, 2019) emphasizes ridge-type regularization in clinical prediction modeling for improving calibration and generalization. Ridge regression is used to: improve risk prediction models, stabilize coefficient estimates in clinical data, support decision-making in healthcare systems.

Text data is high-dimensional and sparse, but ridge regression is often used when dense representations (such as embeddings or tf-idf vectors) are employed. Manning *et al.* (Schütze *et al.*, 2008) describe ridge-type models as effective baselines in information retrieval and text classification tasks. Ridge regression therefore handles large vocabularies efficiently, provides stable classification boundaries, serves as a baseline for more complex NLP models.

Unlike lasso-type methods, ridge regression does not produce sparse solutions but instead distributes shrinkage smoothly across all coefficients, making it particularly suitable for dense, structured signal settings.

Some limitations of ridge regression is that because the ℓ_2 penalty is smooth and strictly convex, the solution shrinks all coefficients toward zero but almost never sets any exactly to zero, so ridge cannot perform automatic feature selection (Hoerl and Kennard, 1970). Also, every coefficient is shrunk by roughly the same proportional factor regardless of its true magnitude, so large signal coefficients are shrunk almost as much as small noise coefficients (Tibshirani, 1996). If the true goal is a parsimonious, interpretable model (*e.g.*, "which 5 of these 200 genes matter"), ridge is the wrong tool — ridge regression is stuck manually thresholding small coefficients after the fact, which has no principled justification and no accompanying uncertainty quantification. In applications where model size matters for deployment cost (*e.g.*, fewer features to collect/measure in a clinical or industrial setting), ridge doesn't help cut anything.

In summary, across applications, ridge regression is most effective when predictors are highly correlated, signals are dense rather than sparse, prediction stability is more important than variable selection.

3. The lasso

Tibshirani (Tibshirani, 1996) introduced the lasso (least absolute shrinkage and selection operator) model:

$$\hat{\beta} = \arg \min \|y - X\beta\|^2 + \lambda \sum |\beta_j|.$$

What was new was that the coefficients can be exactly zero, and the model performs variable selection automatically. A geometric insight is that the ℓ_1 constraint has corners. So the solutions tend to lie on axes, leading to sparsity. From a Bayesian viewpoint the lasso corresponds to:

$$\beta_j \sim \text{Laplace}(0, b).$$

Park and Casella (Park and Casella, 2008) provided a full Bayesian formulation. They showed the Laplace distribution arises as a scale mixture of Normals with an Exponential mixing density. This enables Gibbs sampling, as well as connects shrinkage to hierarchical modeling.

The lasso has been widely adopted across applied fields due to its ability to perform variable selection and regularization simultaneously (Tibshirani, 1996).

Applications and limitations of the lasso

High-throughput genomic experiments produce datasets with thousands of predictors (genes) and relatively few samples. This setting is particularly suited to sparse modeling.

$$\hat{\beta} = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}. \quad (5)$$

The lasso has been used for gene selection in cancer classification problems, where only a small subset of genes is relevant for distinguishing disease subtypes (Golub *et al.*, 1999; Shevade *et al.*, 2000). This identifies small gene panels from thousands of candidates; improves classification accuracy in high-dimensional settings; enables biological interpretability.

Financial prediction problems often involve a large number of correlated predictors and weak signals. The lasso has been used for forecasting asset returns and selecting relevant predictors in macro-finance models (Fan *et al.*, 2008; Belloni and Chernozhukov, 2011). The lasso can reduce overfitting in noisy data; selects economically meaningful predictors; improves out-of-sample forecasting performance. In Healthcare and Clinical Prediction, electronic health records generate high-dimensional datasets with heterogeneous variables.

The lasso is also used to construct clinical risk prediction models, including disease diagnosis and hospital readmission prediction (Tibshirani, 1996; Wu *et al.*, 2009) where it produces sparse and interpretable models; identifies key clinical risk factors; facilitates deployment in clinical settings.

In neuroscience and brain imaging, functional neuro-imaging datasets contain tens of thousands of predictors (voxels) and relatively few observations. Sparse regression methods,

including the lasso, have been used for decoding cognitive states and identifying relevant brain regions (Yamashita *et al.*, 2008; Ryali *et al.*, 2010). This enables identification of task-relevant brain regions, produces interpretable neural decoding models as well as handles extreme high-dimensionality.

In marketing and customer analytics, modern marketing datasets include detailed behavioral and demographic features. Here, The lasso has been used for customer segmentation, churn prediction, and targeted marketing (Verbeke *et al.*, 2012; Netzer *et al.*, 2012). This identifies key drivers of customer behavior, improves targeting efficiency, reduces model complexity.

In Engineering and Fault Detection, industrial systems generate large volumes of sensor data. Here, the lasso has been used in fault detection and system monitoring by identifying relevant signals among many sensors (Wang *et al.*, 2011; Qin and Goldfarb, 2012). This can detect faults early, improves robustness to noise; enhances system reliability.

Text data leads to extremely high-dimensional feature spaces. The lasso has been used for document classification and sentiment analysis through sparse feature selection (Zhang, 2010; Scott and Varian, 2015) as it Selects informative terms; reduces computational cost; improves generalization performance.

The lasso is also used in signal Processing and compressed sensing. Compressed sensing exploits sparsity to recover signals from limited measurements. The lasso formulation has been widely used for signal recovery in compressed sensing (Candès *et al.*, 2006; Donoho, 2006). This Enables accurate reconstruction from few observations; exploits sparsity in signal representation as well as has applications in imaging and communications.

An up-to-date review of Lasso, and Ridge Regression (Gillariose *et al.*, 2026) covers the adaptive Lasso, and open research directions.

Practical trade-off of lasso

Trade-off 1) Irrepresentable Condition Failure — Sparsity vs. Reliability Under Correlation

The practical advantage of using Lasso, is that when predictors are roughly independent (or the irrepresentable condition genuinely holds), Lasso gives exact support recovery — a clean, interpretable subset of “selected” variables with no need for post hoc thresholding. Also, it is computationally cheap and scalable; a single convex problem with fast solvers (coordinate descent, LARS).

A disadvantage is that in practice — genomics, finance, most observational data — predictors are correlated, so the irrepresentable condition is routinely violated (Zhao and Yu, 2006). Lasso’s selected set can then be unstable: small perturbations of the data (bootstrap resamples, different training folds) can flip which variables are selected, even when predictive performance barely changes. Secondly, among a cluster of correlated predictors, Lasso tends to arbitrarily pick one and zero out the rest rather than distributing credit across the group, so “variable X was selected” does not necessarily mean X is causally special — it may simply be the marginal winner within its correlation cluster. This is a genuine problem for scientific inference (*e.g.*, “which gene matters”) but is often less consequential for pure prediction,

since the selected proxy variable still captures the correlated signal.

Practical rule of thumb: do not over-interpret *which* variables Lasso selects when predictors are correlated — trust it for prediction, be skeptical of it for causal or scientific claims about individual variables. Use stability selection (Meinshausen and Bühlmann, 2010) or the Elastic Net if more reliable selection under correlation is required.

Trade-off 2) Constant Shrinkage — Simplicity vs. Accuracy on Strong Effects

Another advantage of Lasso, is that a single tuning parameter λ controls both sparsity and shrinkage simultaneously, making cross-validation simple (a one-dimensional search) and the procedure easy to tune and reason about. Further, the Lasso has a predictable, monotonic behavior, increasing λ always shrinks or removes more variables.

A disadvantage of Lasso, is that every retained coefficient is pulled toward zero by the same amount λ , regardless of whether it corresponds to a modest or a large true effect. This systematically underestimates large true coefficients (Fan and Li, 2001), which matters whenever the downstream use case cares about effect size and not merely which variables are selected. This bias does not vanish unless $\lambda \rightarrow 0$, but $\lambda \rightarrow 0$ also removes the sparsity and regularization benefit — so there is an unavoidable tension: vanilla Lasso cannot simultaneously deliver sparse selection and unbiased large-effect estimates. This is exactly the gap that non-convex penalties (SCAD, MCP) and adaptive/reweighted Lasso variants were designed to close, applying less shrinkage to coefficients already estimated to be large (Zou, 2006).

Practical rule of thumb: If calibrated effect-size estimates are needed (*e.g.*, for reporting inferential confidence, not just prediction), consider a two-stage approach — use Lasso for selection, then refit OLS on the selected subset (“post-Lasso”), or use SCAD/MCP/adaptive Lasso directly (Belloni and Chernozhukov, 2013; Zou, 2006).

Recent work extends the Lasso beyond point estimation and selection toward valid statistical inference in high dimensions.

Debiased (deparsified) lasso

The vanilla Lasso is biased and lacks a tractable asymptotic distribution, so it cannot directly yield confidence intervals. The debiased Lasso corrects this via a bias-correction term based on an approximate inverse of the Gram matrix, giving asymptotically normal estimators for individual coefficients (Zhang and Zhang, 2014; van de Geer *et al.*, 2014). Recent extensions include: 1) Sample-splitting for GLMs: Lasso selects variables on one subsample; debiased Lasso performs inference on a held-out subsample, reducing bias and variance relative to standard MLE (Vazquez *et al.*, 2025). 2) Hidden confounding: the doubly debiased Lasso extends debiasing to settings with unmeasured confounders (Guo *et al.*, 2022). 3) Streaming data: debiasing combined with stochastic gradient descent enables incremental inference without storing the full dataset.

Stability under resampling

Modern theoretical work studies the stability of debiased estimators under data perturbation, linking inferential reliability to resampling-based selection frameworks such as knockoffs (Barber and Candès, 2015) – directly relevant to Lasso’s known instability under correlated designs.

4. Scale mixtures

Andrews and Mallows (Andrews and Mallows, 1974) introduced scale mixtures of normals. The idea is to allow the parameter depend on the shrinkage parameter through the variance;

$$\beta_j | \lambda_j \sim N(0, \lambda_j^2).$$

Here, λ_j controls shrinkage, and randomizing λ_j creates flexibility. For example, if $\lambda_j^2 \sim \text{Inverse-Gamma}$: $\Rightarrow \beta_j \sim t$.

Heavy tails emerge naturally. For example, Spike-and-Slab (Mitchell & Beauchamp, 1988)’s model given by

$$\beta_j \sim (1 - \pi)\delta_0 + \pi N(0, \sigma^2).$$

This model solved the problem of explicit variable selection, interpretable inclusion probabilities, though the limitations are that the model was computationally expensive, required combinatorial model space. To address limitations, hierarchical priors were introduced as mentioned in the next section.

Applications of scale mixtures

Scale mixtures of normal distributions form a flexible and widely used class of shrinkage priors in Bayesian statistics. By representing regression coefficients as:

$$\beta_j \mid \lambda_j \sim \mathcal{N}(0, \lambda_j^2), \quad \lambda_j \sim p(\lambda), \quad (6)$$

these models induce heavy-tailed, marginal priors while retaining conditional conjugacy. This structure enables adaptive shrinkage and efficient computation in high-dimensional problems.

Scale-mixture representations underpin many modern shrinkage priors, including Student- t , Bayesian lasso, Horseshoe, and related global-local priors (Andrews and Mallows, 1974; Park and Casella, 2008; Carvalho *et al.*, 2010). This section highlights key applied domains where scale-mixture-based shrinkage priors have been successfully used.

Genomic datasets involve thousands to millions of genetic markers with relatively small sample sizes. Scale-mixture priors are particularly useful because they allow strong shrinkage of noise variables while preserving signals with large effects.

Carvalho *et al.* (Carvalho *et al.*, 2010) demonstrated that Horseshoe-type priors, which are scale mixtures with heavy-tailed local scales, perform well in sparse genetic signal recovery. Bhattacharya *et al.* (Bhattacharya *et al.*, 2015a) further developed Dirichlet-Laplace priors, another scale-mixture construction, for optimal shrinkage in high-dimensional

genomic settings. The practical outcomes are identification of relevant genetic variants in genome-wide association studies; improved prediction of polygenic traits; effective handling of ultra-high-dimensional marker spaces.

Functional neuroimaging data, such as fMRI, involve extremely high-dimensional voxel-level predictors. Scale-mixture shrinkage priors enable sparse and interpretable brain activity models. Yamashita *et al.* (Yamashita *et al.*, 2008) showed that sparse Bayesian regression methods, closely related to scale mixtures, can identify task-relevant brain regions. Ryali *et al.* (Ryali *et al.*, 2010) applied sparse logistic regression models for whole-brain classification, where scale-mixture priors improve robustness and interpretability. Practical outcome include identification of spatially localized brain activity patterns; improved decoding of cognitive states from neural data; reduction of dimensionality in voxel-based regression models.

Financial datasets often contain large sets of correlated predictors with weak signals. Scale-mixture priors provide robustness through heavy-tailed behavior, allowing extreme values to remain non-regularized. Griffin and Brown (Griffin and Brown, 2010a) introduced normal-gamma scale-mixture priors for regression, which have been applied in financial modeling contexts with many predictors. These priors improve stability in portfolio estimation and factor modeling problems along with improved prediction of asset returns in high-dimensional factor models; stabilization of regression coefficients under multicollinearity; better out-of-sample portfolio performance.

Compressed sensing relies on recovering sparse signals from incomplete observations. Scale-mixture priors provide a Bayesian foundation for ℓ_1 -type regularization and its generalizations. Candès *et al.* (Candès *et al.*, 2006) and Donoho (Donoho, 2006) established the theoretical basis for sparse signal recovery. Bayesian formulations using scale mixtures, including Student- t and Horseshoe priors, have been widely used for robust reconstruction in noisy environments giving accurate signal reconstruction from under-sampled measurements; robustness to measurement noise. This prior is also useful for its applications in MRI and imaging systems. In modern machine learning, scale-mixture priors are widely used for regularized regression and classification problems. Their ability to interpolate between dense and sparse regimes makes them particularly effective. Polson and Scott (Polson and Scott, 2012) developed a general global-local framework based on scale mixtures, which unifies many shrinkage priors used in predictive modeling. These methods are commonly applied in Bayesian versions of logistic regression and hierarchical models. These models can improve predictive performance in high-dimensional classification tasks; adaptively regularize without manual tuning of sparsity levels; offer efficient posterior computation via conjugate mixtures. Text data produces extremely high-dimensional feature spaces, where scale-mixture priors help balance sparsity and flexibility. Scale-mixture-based Bayesian models have been used for document classification and sentiment analysis, where heavy-tailed priors allow a small subset of informative words to dominate predictions while shrinking irrelevant vocabulary features. Sparse selection of informative words or n -grams provide improved generalization in text classification tasks as well as reduce overfitting in high-dimensional bag-of-words models.

Practical trade-offs of scale-mixture shrinkage priors

1. Computation: scale mixtures avoid the combinatorial search of spike-and-slab priors (Polson and Scott, 2010), but naive Gibbs/HMC samplers scale poorly in high dimensions, requiring specialized fast samplers (Johndrow *et al.*, 2020). 2. Flexibility vs. tuning: the hierarchy $\beta_j \mid \psi_j, \tau \sim N(0, \psi_j \tau)$ unifies the Bayesian Lasso, Horseshoe, and Dirichlet–Laplace priors (Carvalho *et al.*, 2010; Bhattacharya *et al.*, 2015b), but every instance still needs a global scale hyperparameter, and standard defaults are largely conventional rather than principled (Griffin and Brown, 2010b). 3. Structured extensions: the framework extends to dependent data via product representations (Griffin and Hoff, 2024) or spatial sandwich-mixture priors (Nishina *et al.*, 2026), the latter paper builds directly on an earlier, (non-spatial) global-local small-area framework (Tang *et al.*, 2018) but each extension introduces new identifiability and posterior propriety issues to verify. 4. No exact sparsity: these priors concentrate well near sparse vectors, with Dirichlet–Laplace outperforming the Bayesian Lasso in simulation (Bhattacharya *et al.*, 2015b), but continuous mixing means no posterior draw is ever exactly zero, so “selection” still requires post hoc thresholding (Carvalho *et al.*, 2010).

Scenario	Consequence	Mitigation
Very high-dimensional p	Slow mixing	Fast samplers (Johndrow <i>et al.</i> , 2020)
Choosing global scale	Manual/hyperprior tuning	Well-studied defaults
Correlated/spatial predictors	Dependence ignored	Structured priors (Griffin and Hoff, 2024; Nishina <i>et al.</i> , 2026)
Need exact sparse output	No exact zeros	Post hoc thresholding

5. The horseshoe prior

Another noteworthy shrinkage priors is the Horseshoe prior given by the following:

$$\beta_j \mid \lambda_j, \tau \sim \mathcal{N}(0, \tau^2 \lambda_j^2), \quad (7)$$

$$\lambda_j \sim \text{Cauchy}^+(0, 1), \quad (8)$$

$$\tau \sim \text{Cauchy}^+(0, 1). \quad (9)$$

This prior has two key properties:

1. an infinite spike at zero, encouraging strong shrinkage,
2. heavy tails, allowing large signals to escape shrinkage.

These features make it particularly effective in sparse settings. Later work refined its interpretation and hyperparameter specification (Piironen and Vehtari, 2017a). This is a

continuous analogue of spike-and-slab, namely the Horseshoe Prior (Carvalho *et al.*, 2010) who showed this prior strongly shrinks noise, preserves large signals, behaves like spike-and-slab without discreteness. For example, suppose there are 100 coefficients, 95 are near zero, 5 are large. Then: small coefficients $\rightarrow$ heavily shrunk, large ones $\rightarrow$ almost unchanged. This works because the prior has infinite mass near zero, as well as very heavy tails.

Connection of horseshoe priors to spike-and-slab priors

Earlier Bayesian variable selection methods used spike-and-slab priors:

$$\beta_j \sim (1 - \pi)\delta_0 + \pi\mathcal{N}(0, \sigma^2), \quad (10)$$

which explicitly model inclusion/exclusion.

Continuous shrinkage priors such as the Horseshoe can be seen as smooth relaxations of spike-and-slab models, avoiding discrete model selection.

Practical trade-off of the horseshoe prior

Strong sparsity vs. unregularized tails: the half-Cauchy local scales give near-spike behavior at zero, but the same heavy tails leave large signals essentially unregularized (Carvalho *et al.*, 2010). This is problematic under weak identifiability (*e.g.* separation in logistic regression), where the likelihood alone cannot constrain large coefficients, causing instability or divergence in MCMC (Piironen and Vehtari, 2017b). The regularized Horseshoe fixes this by capping prior variance via a finite slab scale, trading a small amount of tail flexibility for bounded, stable inference (Piironen and Vehtari, 2017b).

No separate control of sparsity vs. large-coefficient regularization: the standard Horseshoe cannot specify sparsity belief and the regularization of large coefficients independently, which recent robustness studies confirm can still cause numeric instability in MCMC when relying purely on infinite-variance half-Cauchy priors (Fan *et al.*, 2026). The regularized Horseshoe addresses this directly by decoupling the two via an explicit slab width parameter (Piironen and Vehtari, 2017b).

Computational cost at scale: per-iteration MCMC cost for the Horseshoe has historically hindered its use on large-scale data; recent software addresses this with approximate/adaptive algorithms that exploit an effective number of nonzero coefficients to skip negligible ones (Kang and Lee, 2025). The following table gives a summary comparison across models

Method	Sparsity	Bias (large signals)	Adaptivity
Ridge	No	High	Low
Lasso	Yes	Moderate	Medium
Spike-slab	Yes	Low	High
Horseshoe	Yes	Very low	Very high

6. Theoretical properties of shrinkage priors

A central question in high-dimensional Bayesian inference is whether posterior distributions concentrate around the true parameter at an optimal rate.

Posterior contraction

Let β_0 denote the true parameter vector, assumed sparse with s_0 nonzero entries.

Theorem 1 (Posterior Contraction for Sparse Signals): Consider the Gaussian sequence model:

$$Y_j = \beta_{0j} + \varepsilon_j, \quad \varepsilon_j \sim N(0, 1).$$

Under a global-local shrinkage prior with sufficiently heavy tails (*e.g.*, Horseshoe), the posterior satisfies:

$$\Pi \left(\|\beta - \beta_0\|_2 \leq M \sqrt{s_0 \log p} \mid Y \right) \rightarrow 1,$$

for some constant $M > 0$ as $p \rightarrow \infty$.

This can be interpreted as the following: the posterior concentrates at (near) the minimax rate for sparse estimation. The argument of the proof proceeds in three steps:

1. **Prior mass near truth:** Heavy-tailed priors assign sufficient probability to neighborhoods of large signals.
2. **Testing argument:** Construct tests separating β_0 from alternatives at distance ϵ_n .
3. **Control of small coefficients:** The prior places strong mass near zero, shrinking noise coordinates.

Remark 1: The heavy tails of the prior are essential: light-tailed priors (*e.g.*, Gaussian) fail to allocate enough mass to large signals, leading to suboptimal contraction.

More refined results show that the Horseshoe prior adapts to unknown sparsity levels and achieves near-optimal risk without explicit tuning.

Polson and Scott (Polson and Scott, 2010, 2012) developed a general framework for global-local shrinkage priors, which unifies a wide range of Bayesian regularization methods including the Horseshoe, Bayesian lasso, and normal-gamma priors.

Shrinkage factor representation

A key insight of this framework is that posterior shrinkage can be expressed through latent shrinkage factors. Define:

$$\kappa_j = \frac{1}{1 + \tau^2 \lambda_j^2},$$

where τ is a global scale parameter and λ_j is a local scale parameter associated with coefficient β_j .

Under Gaussian likelihood models, posterior means admit the approximate form:

$$\mathbb{E}[\beta_j \mid Y_j] \approx (1 - \kappa_j) Y_j.$$

This representation provides a direct interpretation of shrinkage behavior. Figure 3 demonstrates this in the Appendix:

1. $\kappa_j \approx 1$ implies strong shrinkage toward zero (noise-dominated coefficients),
2. $\kappa_j \approx 0$ implies weak shrinkage (signal-dominated coefficients).

7. Simulation comparison: ridge, lasso, and horseshoe

Setup

We compare Ridge (Hoerl and Kennard, 1970), the Lasso (Tibshirani, 1996), and the horseshoe prior (Carvalho *et al.*, 2010) on coefficient estimation error and out-of-sample prediction error under a sparse, correlated design. For each of $B = 30$ replications we generate

$$X_i \sim N(0, \Sigma), \quad \Sigma_{jk} = \rho^{|j-k|}, \quad y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \quad (11)$$

with $n_{\text{train}} = 100$, $n_{\text{test}} = 300$, $p = 100$, Toeplitz correlation $\rho = 0.5$, and a sparse true β with $s_0 = 10$ nonzero entries (magnitude 1.5, random sign and location). The noise variance σ^2 is set so that $\text{Var}(X\beta)/\sigma^2 = 4$ (SNR = 4). All three priors are fit by Gibbs sampling under the same MCMC settings (1500 iterations, 700 burn-in, posterior mean reported), so the comparison isolates the effect of the *prior*, not the fitting procedure.

Models

$$\text{Bayesian Ridge: } \beta_j \mid \tau^2, \sigma^2 \sim N(0, \sigma^2 \tau^2), \quad \tau^2 \sim \text{IG}(\epsilon, \epsilon) \quad (12)$$

$$\text{Bayesian Lasso: } \beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2 \tau_j^2), \quad \tau_j^2 \sim \text{Exp}(\lambda^2/2), \quad \lambda^2 \sim \text{Gamma}(1, 1) \quad (13)$$

$$\text{Horseshoe: } \beta_j \mid \lambda_j, \tau, \sigma^2 \sim N(0, \sigma^2 \tau^2 \lambda_j^2), \quad \lambda_j, \tau \sim C^+(0, 1) \quad (14)$$

The Bayesian Ridge shares a single global variance τ^2 across all coefficients (Gibbs updates: Normal for β , Inverse-Gamma for σ^2 and τ^2). The Bayesian Lasso follows Park and Casella (2008), with $1/\tau_j^2$ updated via an Inverse-Gaussian full conditional. The horseshoe uses the Makalic and Schmidt (2016) parameter expansion. Data are generated: $n_{\text{train}} = 100$, $p = 100$, Toeplitz correlation $\rho = 0.5$, $s_0 = 10$ true nonzero coefficients, SNR = 4, $B = 30$ replications.

Results

Method	Coef. MSE	Pred. MSE	Bias (signal)	Bias (null)
Ridge (Gibbs)	0.0979 ± 0.0024	12.55 ± 0.28	0.0228 ± 0.0384	-0.0037 ± 0.0044
Lasso (Gibbs)	0.0982 ± 0.0044	12.89 ± 0.39	0.0288 ± 0.0242	-0.0049 ± 0.0040
Horseshoe (Gibbs)	0.0243 ± 0.0026	7.70 ± 0.31	0.0397 ± 0.0202	-0.0032 ± 0.0020

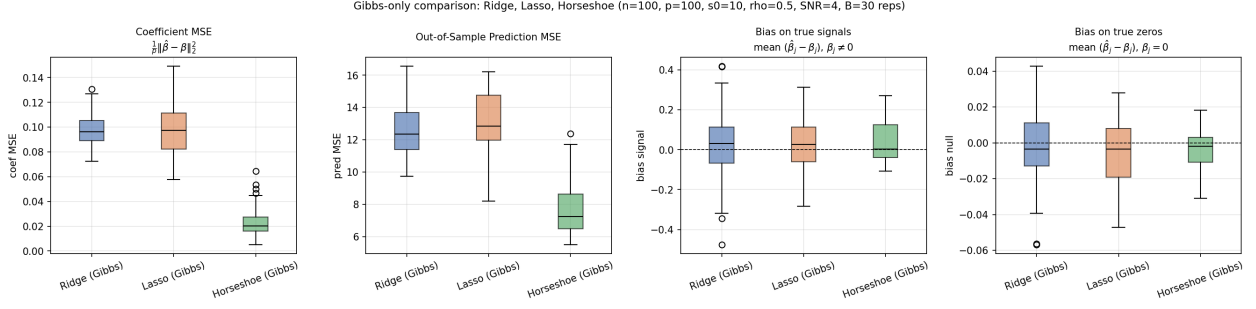


Figure 1: Coefficient MSE, prediction MSE, and bias (on true signal and true null coefficients) across $B = 30$ replications, for the Gibbs-sampled Ridge, Lasso, and horseshoe posteriors.

Discussion: bias vs. variance

Bias on the true null coefficients is negligible and statistically indistinguishable across all three priors (≈ -0.003 to -0.005 , well within one standard error of one another): every method correctly shrinks noise coefficients toward zero.

The comparison on true signal coefficients is more subtle. The horseshoe does *not* have the smallest point estimate of bias – at 0.0397 it is in fact the largest of the three – yet it achieves the lowest coefficient MSE. This is reconciled by the classical decomposition

$$\text{MSE} = \text{Bias}^2 + \text{Variance}, \quad (15)$$

together with the horseshoe’s markedly smaller standard error on the bias estimate itself (0.0202, roughly half that of Ridge’s 0.0384): its posterior-mean estimates are more stable across replications, even though their average deviation from the true signal is not smaller. The horseshoe’s MSE advantage over Ridge and Lasso (Gibbs) in this experiment is therefore driven primarily by *variance reduction*, not by reduced bias.

Ridge and Lasso (Gibbs) show similar bias magnitudes to each other (0.023 vs. 0.029) and similarly wide standard errors, consistent with the earlier finding that the weakly-tuned Bayesian Lasso hyperprior on λ^2 causes it to behave more like Ridge than like a sparsity-inducing estimator in this setting.

Given the overlapping standard errors on all three bias estimates, none of the pairwise bias differences would be considered statistically distinguishable at $B = 30$ replications; a larger number of replications or a paired test across replications would be needed before concluding that any one prior has systematically lower bias than another.

8. Conclusion

In summary, we may view shrinkage as the “bias” introduced to reduce variance. The theory introduces heavy tailed distributions as a means of protection for large signals. Some methods introduce hierarchy to include adaptive behavior as in Bayesian hierarchical models. Sometimes continuous priors are introduced to approximate discrete selection.

The evolution of shrinkage priors reflects a progression. First, Stein proposed Ridge regression. Then came the Lasso, followed by Bayesian Hierarchies and finally Global-Local

Priors. Each step addressed a limitation of the previous approach: ridge regression addressed stability, lasso: sparsity, Bayesian methods: uncertainty, horseshoe: adaptivity. Bayesian shrinkage priors now form a central tool in high-dimensional inference, balancing flexibility, interpretability, and theoretical guarantees. Shrinkage priors, also, popularly known as regularization priors, unify ideas from decision theory, optimization, and Bayesian inference. Their ability to adaptively regularize high-dimensional models makes them indispensable in modern statistics as well as complex settings.

References

- Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B*, **36**, 99–102.
- Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, **43**, 2055–2085.
- Belloni, A. and Chernozhukov, V. (2011). High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation: Stats in the Château Summer School, August 31-September 4, 2009*, pages 121–156. Springer.
- Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–547.
- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015a). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.
- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015b). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.
- Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, **52**, 489–509.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.
- De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, **182**, 375–385.
- DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, **22**, 1915–1953.
- Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, **52**, 1289–1306.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.
- Fan, J., Zhang, J., and Yu, K. (2008). Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *arXiv preprint arXiv:0812.2604*, .
- Fan, K., Subedi, S., Dissanayake Pathirana, V. R., and Wu, C. (2026). Robust Bayesian high-dimensional variable selection and inference with the horseshoe family of priors. *Computational Statistics & Data Analysis*, **219**, 108358.

- Gillariose, J., Joseph, J., and Chesneau, C. (2026). Lasso and ridge regression: A comprehensive review of applications and developments in machine learning. *International Journal of Data Science and Analytics*, **21**, 7.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Griffin, J. E. and Brown, P. J. (2010a). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.
- Griffin, J. E. and Brown, P. J. (2010b). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.
- Griffin, M. and Hoff, P. D. (2024). Structured shrinkage priors. *Journal of Computational and Graphical Statistics*, **33**.
- Guo, Z., Čevd, D., and Bühlmann, P. (2022). Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, **50**, 1320–1347.
- Hansen, P. C. (1998). *Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion*. SIAM.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
- Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios. *Journal of Finance*, **4**, 1651–83.
- Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, **21**, 1–61.
- Kang, M. and Lee, K. (2025). Mhorseshoe package in R: Approximate algorithm for the horseshoe prior in Bayesian linear model. *SoftwareX*, **31**, 102236.
- Mann, M. E., Bradley, R. S., and Hughes, M. K. (1998). Global-scale temperature patterns. *Nature*, **392**, 779–787.
- Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, **72**, 417–473.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.
- Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, **31**, 521–543.
- Nishina, S., Onizuka, T., and Hashimoto, S. (2026). Global-local shrinkage priors for modeling random effects in multivariate spatial small area estimation. *arXiv preprint arXiv:2601.14752*, .
- Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.
- Piironen, J. and Vehtari, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *Electronic Journal of Statistics*, **11**, 5018–5051.
- Piironen, J. and Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, **11**, 5018–5051.

- Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, **9**, 501–538.
- Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**, 887–902.
- Qin, Z. T. and Goldfarb, D. (2012). Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, **13**.
- Ryali, S. et al. (2010). Sparse logistic regression for fMRI classification. *NeuroImage*, **51**.
- Saunders, C. et al. (1998). Ridge regression learning algorithm in dual variables. *ICML*, , 515–521.
- Schneider, T. (2001). Analysis of incomplete climate data. *Journal of Climate*, **14**, 853–871.
- Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*, volume 39. Cambridge University Press Cambridge.
- Scott, S. L. and Varian, H. R. (2015). Bayesian variable selection for nowcasting economic time series. In *Economic Analysis of the Digital Economy*, pages 119–135. University of Chicago Press.
- Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, **11**, 1188–1193.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206.
- Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Springer, Cham, Switzerland, 2 edition.
- Tang, X., Ghosh, M., Ha, N. S., and Sedransk, J. (2018). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, **113**, 1476–1489.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267–288.
- Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. Winston distributed solely by Halsted Press.
- van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.
- Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *U.C. Berkeley Division of Biostatistics Working Paper Series*, **222**.
- Vazquez, D. L. et al. (2025). Debiased lasso after sample splitting for estimation and inference in high-dimensional generalized linear models. *Canadian Journal of Statistics*, **53**, e11827.
- Verbeke, M., Morante, R., Frasconi, P., and De Raedt, L. (2012). A statistical relational learning approach to identify sections in scientific abstracts using sentence and document structure. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*.

- Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*, **5**, 468.
- Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.
- Yamashita, O. et al. (2008). Sparse estimation for fMRI decoding. *NeuroImage*, **42**, 1414–1429.
- Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, **76**, 217–242.
- Zhang, J. (2010). The influence machine: Nonnegative instance-space learning with differentiated regularization. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 861–866. IEEE.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.