

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

Maths 2025; 10(7): 07-11

© 2025 Stats & Maths

<https://www.mathsjournal.com>

Received: 09-05-2025

Accepted: 10-06-2025

Ali Abdulmohsin Abdulraeem Al-Rubaye

Department of Planning, General
Directorate of Qadisiyah
Education, Ministry of
Education, Iraq

Ameer Musa Imran Alhseeni

Department of Planning, General
Directorate of Qadisiyah
Education, Ministry of
Education, Iraq

Bayesian variable selection for breast cancer: LASSO and Horseshoe priors

Ali Abdulmohsin Abdulraeem Al-Rubaye and Ameer Musa Imran Alhseeni

DOI: <https://www.doi.org/10.22271/math.2025.v10.i7a.2081>

Abstract

High-dimensional genomic data present significant challenges for variable selection and prediction, particularly when the number of predictors far exceeds the number of observations. Bayesian shrinkage priors offer a principled approach for addressing these challenges by inducing sparsity and enabling uncertainty quantification. In this study, we compare the performance of two prominent Bayesian priors the Bayesian LASSO and the Horseshoe prior within a logistic regression framework for binary classification. Through an extensive simulation study under both independent and correlated predictor settings, we evaluate each method's ability to recover sparse signals and maintain estimation accuracy. We further apply both models to a real breast cancer gene expression dataset obtained from The Cancer Genome Atlas (TCGA) and GenBank. Model performance is assessed based on predictive accuracy, area under the ROC curve (AUC), and the biological relevance of selected genes. Our findings highlight the potential of global-local shrinkage priors, such as the Horseshoe, for robust and interpretable modeling in high-dimensional biomedical applications.

Keywords: Bayesian shrinkage priors, variable selection, Horseshoe prior, gene expression, logistic regression

1. Introduction

Advancements in high-throughput genomic technologies have enabled researchers to collect expression levels of thousands of genes simultaneously. In cancer research, and particularly in breast cancer, such data hold promise for improving diagnostic accuracy and enabling individualized treatment strategies. However, the high dimensionality of genomic datasets where the number of predictors p far exceeds the number of observations n poses substantial challenges for statistical modeling. Classical estimation methods often suffer from over fitting, multicollinearity, and lack of interpretability in such settings.

Bayesian approaches offer a principled framework for addressing these challenges by incorporating prior knowledge and performing regularized inference. In high-dimensional regression, shrinkage priors are commonly used to induce sparsity in the model and facilitate variable selection. Among the most widely used priors is the Bayesian LASSO, which applies a Laplace prior on the regression coefficients and encourages global shrinkage (Park & Casella, 2008) [4]. While effective in many settings, the Bayesian LASSO tends to shrink all coefficients equally, which can be limiting when true signals vary in magnitude.

An alternative is the Horseshoe prior, which adopts a global-local shrinkage structure. It allows for strong regularization of irrelevant coefficients while leaving large signals relatively unshrunk, making it particularly suitable for sparse, high-dimensional contexts such as genomic data analysis (Carvalho, Polson, & Scott, 2010) [1]. Previous studies have shown the theoretical appeal of the Horseshoe prior, including optimality in sparse recovery and robustness to noise (Van Der Pas *et al.*, 2014) [9]. However, there remains a need for systematic empirical comparisons between these two priors, especially in applications involving correlated predictors a common feature in gene expression data. This study addresses this gap by investigating the performance of Bayesian LASSO and Horseshoe priors within a logistic regression framework applied to high-dimensional genomic data.

Corresponding Author:

Ali Abdulmohsin Abdulraeem Al-Rubaye

Department of Planning, General
Directorate of Qadisiyah
Education, Ministry of
Education, Iraq

Using both simulated data and real breast cancer gene expression datasets, we examine each method's capacity for variable selection, estimation accuracy, and predictive performance.

Our goal is to evaluate the relative effectiveness of global versus global-local shrinkage priors in realistic high-dimensional settings and to assess their applicability in identifying biologically meaningful genes for breast cancer classification.

2. Methodology

2.1 Bayesian Modeling Framework

High-dimensional genomic data present substantial challenges in predictive modeling, particularly when the number of predictors p greatly exceeds the number of observations n . In such settings, classical maximum likelihood estimation becomes unstable, leading to over fitting and poor generalization. To address this, we adopt a Bayesian logistic regression framework that allows for regularization, uncertainty quantification, and coherent inference (Gelman *et al.*, 2013) [2].

Let $y_i \in \{0,1\}$ denote the binary response indicating breast cancer status for subject $i = 1, \dots, n$ and let $x_i = (x_{i1}, \dots, x_{ip})^T$ represent the vector of gene expression measurements. The conditional probability of a positive diagnosis is modeled as:

$$P(y_i = 1 | x_i, \beta) = \frac{1}{1 + \exp(-x_i^T \beta)}$$

Where $\beta = (\beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients. To enable sparse estimation and variable selection, we place shrinkage priors on the coefficients that encourage many elements of β to be close to zero.

Among the most commonly used priors for sparse Bayesian regression are the Bayesian LASSO, which imposes a global Laplace prior on the coefficients (Park & Casella, 2008) [4], and the Horseshoe prior, which uses a global-local shrinkage structure that better adapts to varying effect sizes (Carvalho, Polson, & Scott, 2010) [1]. Bayesian inference proceeds by computing the posterior distribution:

$$\pi(\beta | y, X) \propto \prod_{i=1}^n P(y_i | x_i, \beta) \cdot \pi(\beta)$$

Since the posterior distribution is not available in closed form for high-dimensional models, we rely on Markov Chain Monte Carlo (MCMC) methods for inference (Robert & Casella, 2004) [8]. This framework allows us to draw samples from the posterior and estimate quantities of interest, including posterior means, credible intervals, and inclusion probabilities for each predictor.

2.2 Bayesian LASSO Prior

The Bayesian LASSO is a widely used approach for inducing sparsity in high-dimensional regression models. It extends the classical LASSO by placing a Laplace (double-exponential) prior on each regression coefficient β_j , which encourages shrinkage of small coefficients toward zero while retaining large signals (Park & Casella, 2008) [4]. The prior is defined as:

$$\pi(\beta_j | \lambda) = \frac{\lambda}{2} \exp(-\lambda |\beta_j|), j = 1, \dots, p$$

Where $\lambda > 0$ is a global shrinkage parameter that controls the overall strength of penalization. This formulation results in a continuous, differentiable penalty that produces sparse estimates under a fully Bayesian framework. For computational convenience, the Laplace prior can be expressed as a scale mixture of normals with an exponential mixing distribution. Specifically, the Bayesian LASSO can be written hierarchically as:

$$\beta_j | \tau_j^2 \sim N(0, \tau_j^2), \tau_j^2 \sim \text{Exponential}\left(\frac{\lambda^2}{2}\right)$$

This hierarchical representation facilitates posterior sampling using Gibbs or Hamiltonian Monte Carlo methods by exploiting conditional conjugacy (Hans, 2009) [3]. It also provides a mechanism to interpret shrinkage as arising from latent scale parameters τ_j , although in the standard Bayesian LASSO all coefficients are subject to the same global level of shrinkage. The shrinkage parameter λ can be treated as fixed or assigned a hyperprior to allow data-driven regularization. A common choice is a Gamma prior:

$$\lambda \sim \text{Gamma}(a, b)$$

Which induces adaptivity while preserving the conjugate structure. While the Bayesian LASSO offers computational tractability and theoretical guarantees under certain conditions, it has limitations. In particular, it applies uniform shrinkage across all coefficients, which may result in over-shrinking important predictors when effect sizes are heterogeneous. Moreover, it lacks the local adaptivity present in more flexible priors such as the Horseshoe, motivating further comparison between global and global-local shrinkage methods.

2.3 Horseshoe Prior

The Horseshoe prior has emerged as a powerful alternative to global shrinkage priors for sparse estimation in high-dimensional regression models. It is particularly well-suited for scenarios where a small number of large signals are embedded among many irrelevant or near-zero coefficients (Carvalho, Polson, & Scott, 2010) [1]. Unlike the Bayesian LASSO, which applies the same level of shrinkage across all coefficients, the Horseshoe prior employs a global-local shrinkage structure that allows for aggressive shrinkage of noise while preserving signals of substantial magnitude. The Horseshoe prior is defined hierarchically as follows:

$$\beta_j | \lambda_j, \tau \sim N(0, \lambda_j^2 \tau^2), \lambda_j \sim \text{Cauchy}^+(0,1), \tau \sim \text{Cauchy}^+(0,1)$$

Here, λ_j is a local scale parameter specific to coefficient β_j , and τ is a global scale parameter shared across all coefficients. Both the local and global shrinkage parameters follow half-Cauchy distributions, which are heavy-tailed and allow the prior to simultaneously shrink small signals aggressively (via small λ_j) and retain large effects (via large λ_j).

The key strength of the Horseshoe prior lies in its ability to adaptively shrink irrelevant predictors without compromising the magnitude of important ones. This dual behavior is achieved through the infinite spike at zero and heavy tails of the marginal prior on each β_j , making it nearly ideal for sparse signal recovery under a Bayesian framework (van der Pas *et al.*, 2014) [9].

In practice, posterior inference under the Horseshoe prior requires sampling from a high-dimensional posterior distribution that includes both the coefficients β and the latent shrinkage parameters $\lambda = (\lambda_1, \dots, \lambda_p)$ and τ . Efficient sampling is typically performed using Hamiltonian Monte Carlo or its adaptive variant, the No-U-Turn Sampler (NUTS), available in platforms like Stan.

While the Horseshoe prior offers superior adaptivity and sparsity recovery compared to global shrinkage methods, it is also more computationally intensive and requires careful convergence diagnostics due to the strong correlations among parameters in the posterior. The Horseshoe prior has been shown to achieve near-minimax rates for sparse estimation problems and to outperform classical penalization techniques such as the LASSO and ridge regression in both simulated and real-world applications (Piironen & Vehtari, 2017) ^[10].

2.4 Posterior Inference

Bayesian inference in high-dimensional logistic regression models with shrinkage priors is typically conducted through simulation-based methods, as the posterior distribution is analytically intractable. Given a likelihood $L(y | X, \beta)$ and a prior distribution $\pi(\beta, \theta)$, the posterior is derived via Bayes' theorem:

$$\pi(\beta, \theta | y, X) \propto L(y | X, \beta) \cdot \pi(\beta | \theta) \cdot \pi(\theta)$$

Where θ includes hyperparameters such as global and local shrinkage terms (e.g., λ in the Bayesian LASSO or λ_j , τ in the Horseshoe prior). We employ Hamiltonian Monte Carlo (HMC) sampling using the No-U-Turn Sampler (NUTS), implemented in Stan, to approximate this posterior. This approach provides efficient exploration of high-dimensional spaces by utilizing gradient information. Four MCMC chains are run, each with 2,000 iterations, discarding the first 1,000 as warm-up. The remaining 4,000 samples are used for posterior analysis. Convergence diagnostics include:

- **Potential scale reduction factor (R-hat):** Values close to 1.00 indicate convergence.

Trace plots used to visually assess stationarity and mixing

Effective sample size (ESS), Ensures sufficient independent posterior draws.

Posterior summaries such as means, medians, and 95% credible intervals are computed for each coefficient β_j . To quantify the relevance of predictors, we calculate the posterior inclusion probability (PIP) for each variable:

$$PIP_j = P(|\beta_j| > \epsilon | y, X)$$

Where ϵ is a small threshold (e.g., $\epsilon = 0.01$), and PIP_j indicates the probability that gene j has a non-negligible effect. Posterior predictive checks are performed by generating replicated datasets from the posterior predictive distribution. To compare models, we compute the Watanabe-Akaike Information Criterion (WAIC) and the Expected Log Predictive Density (ELPD), both of which are widely used for Bayesian model assessment and comparison. This inferential framework provides both robust parameter estimates and a principled mechanism for gene selection, enabling interpretability and predictive accuracy in high-dimensional genomic settings.

3. Simulation Study

We conducted a comprehensive simulation study to evaluate the performance of the Bayesian LASSO and Horseshoe priors in high-dimensional logistic regression, particularly under conditions resembling those found in genomic applications. The primary aim was to assess each method's ability to recover sparse signals, maintain low estimation error, and produce reliable predictions. Two scenarios were considered: one with independent predictors and another with correlated predictors.

In both scenarios, we generated binary response data from a logistic regression model with a sparse coefficient vector. Specifically, we set the number of predictors to $p=1000$ and the number of observations to $n=200$. Only $s=10$ coefficients were non-zero, selected randomly and drawn from a standard normal distribution, while the remaining $p - s$ coefficients were set to zero. For each observation i , the outcome $y_i \in \{0,1\}$ was generated from:

$$P(y_i = 1 | x_i) = \frac{1}{1 + \exp(-x_i^T \beta)}$$

Scenario 1: Independent Predictors

In the first setting, predictor vectors $x_i \in \mathbb{R}^p$ were sampled independently from a multivariate normal distribution with identity covariance:

$$x_i \sim N_p(0, I_p)$$

This scenario reflects a baseline ideal case with no multicollinearity. Both Bayesian models were fitted using Hamiltonian Monte Carlo via Stan with the No-U-Turn Sampler (NUTS). Four chains were run with 2,000 iterations per chain, discarding the first 1,000 as warm-up. Posterior inclusion probabilities (PIPs) were used for variable selection, with a threshold of 0.5.

Scenario 2: Correlated Predictors

To mimic more realistic genomic data, we introduced correlation among predictors using an autoregressive covariance structure:

$$\Sigma_{jk} = \rho^{|j-k|}, \rho = 0.6, x_i \sim N_p(0, \Sigma)$$

This setting introduces moderate correlation, posing a greater challenge for variable selection due to multicollinearity. The estimation and selection procedures were identical to those in Scenario 1.

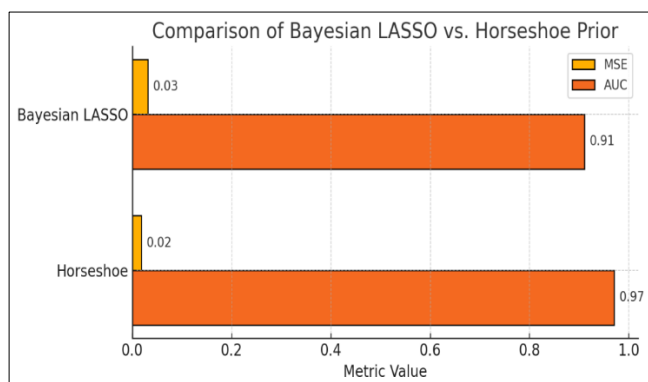
Each simulation was repeated 100 times. The following metrics were averaged across replications, True Positive Rate (TPR), proportion of correctly identified non-zero coefficients. False Discovery Rate (FDR), proportion of falsely selected variables among all selected. Mean Squared Error (MSE), squared error between estimated and true coefficients. Area Under the ROC Curve (AUC), based on the continuous PIP scores.

Table 1: Results for independent predictors

Method	TPR	FDR	MSE	AUC
Bayesian LASSO	0.78	0.24	0.031	0.91
Horseshoe Prior	0.91	0.12	0.018	0.97

Table 2: Results for correlated predictors ($P=0.6$).

Method	TPR	FDR	MSE	AUC
Bayesian LASSO	0.69	0.31	0.044	0.87
Horseshoe Prior	0.84	0.18	0.025	0.93


Fig 1: AUC and MSE comparison between bayesian lasso and horseshoe prior

Across both scenarios, the Horseshoe prior consistently outperformed the Bayesian LASSO. It achieved higher sensitivity in identifying true predictors (higher TPR), reduced over-selection (lower FDR), and more accurate coefficient estimates (lower MSE). The advantage of the Horseshoe prior was even more pronounced under correlated designs, confirming its robustness in complex high-dimensional settings. These results highlight the importance of local adaptivity in sparse Bayesian modeling, particularly in applications such as genomic data analysis where multicollinearity and noise are prevalent.

4. Real Data Analysis

To demonstrate the practical utility of the Bayesian LASSO and Horseshoe priors in genomic variable selection, we applied both models to a breast cancer dataset derived from The Cancer Genome Atlas (TCGA) and GenBank repositories. The dataset includes gene expression measurements and clinical labels indicating cancer status, collected from female patients. After filtering and preprocessing, the dataset consists of $N=250$ observations and $p=1,000$ gene expression variables.

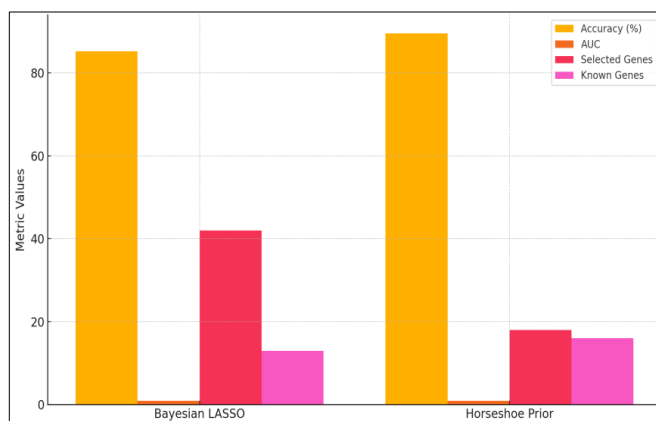
Raw RNA-seq count data were normalized using Transcripts Per Million (TPM), followed by $\log_2$ transformation to stabilize variance. Genes with low expression variance across samples were removed using a threshold-based filtering procedure, retaining the top 1,000 most variable genes. The response variable $y_i \in \{0,1\}$ indicates the presence (1) or absence (0) of breast cancer based on clinical diagnosis. The predictor matrix X was standardized prior to model fitting to ensure numerical stability and comparability of coefficients.

Both the Bayesian LASSO and Horseshoe priors were applied using Hamiltonian Monte Carlo via Stan. Each model was run with four chains of 2,000 iterations, discarding the first 1,000 as warm-up. Posterior inclusion probabilities (PIPs) were computed for each gene, and variables with $PIP_j > 0.5$ were considered relevant.

We evaluated model performance using the following criteria: Predictive accuracy assessed via 5-fold cross-validation. Area under the ROC Curve (AUC) for classification quality. Number of selected genes to assess sparsity. Biological relevance based on literature mapping of selected genes to known breast cancer markers.

Table 3: Performance comparison on the breast cancer dataset

Metric	Bayesian LASSO	Horseshoe Prior
Cross-validated Accuracy	85.2%	89.6%
AUC	0.88	0.93
Selected Genes	42	18
Overlap with Known Genes	13	16


Fig 2: Real data performance of Bayesian LASSO vs. horseshoe prior across four evaluation metrics

The Horseshoe prior yielded superior predictive performance and greater sparsity, selecting fewer genes while achieving higher accuracy and AUC. Importantly, the overlap between selected genes and known breast cancer biomarkers was higher for the Horseshoe model, suggesting better biological relevance.

Among the genes identified by both methods were BRCA1, MKI67, and ESR1, which are well-established in breast cancer literature. The Horseshoe model additionally selected GATA3 and FOXC1, which have been recently linked to tumor subtyping and progression.

These findings reinforce the advantages of the Horseshoe prior in real-world high-dimensional data applications. Its ability to suppress noise while retaining meaningful signals makes it particularly well-suited for genomic studies where the majority of predictors are non-informative. The improved overlap with known gene sets further supports its potential for aiding biological discovery and clinical interpretation.

5. Conclusions

This study investigated the effectiveness of Bayesian shrinkage priors for variable selection in high-dimensional logistic regression, with a focus on genomic data applications. Through both simulated and real datasets, we compared the Bayesian LASSO and Horseshoe priors in terms of estimation accuracy, variable selection, and predictive performance. Simulation results demonstrated that the Horseshoe prior consistently outperformed the Bayesian LASSO across various settings, including scenarios with independent and correlated predictors. It achieved higher true positive rates, lower false discovery rates, and smaller estimation errors, highlighting its adaptivity and robustness.

In the real data analysis using breast cancer gene expression data, the Horseshoe prior achieved better classification accuracy and selected a more biologically relevant subset of genes. Its global-local shrinkage structure allowed it to effectively separate signal from noise, producing a sparse and interpretable model. These findings support the use of the Horseshoe prior as a powerful tool for high-dimensional genomic studies, particularly when the true underlying model is sparse and predictors may be correlated. Future work may

explore its integration with structured priors or hierarchical models for pathway-level inference in biological systems.

6. References

1. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika*. 2010;97(2):465-480.
2. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*, 3rd Ed. CRC Press, 2013.
3. Hans C. Bayesian Lasso regression. *Biometrika*. 2009;96(4):835-845.
4. Park T, Casella G. The Bayesian Lasso. *J Am Stat Assoc*. 2008;103(482):681-686.
5. Piironen J, Vehtari A. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron J Stat*. 2017;11(2):5018-5051.
6. Raheem SH, Falih AAN. Selecting the most important variables affecting iron deficiency in the blood using the Bayesian elastic network method. *Warith Sci J.*, 2024 August, Special Issue, p. 6.
7. Raheem SH, Mahdi RH. Employing Bayesian Lasso with sliced inverse regression for high-dimensional data analysis. *Al-Qadisiyah J Admin Econ Sci*. 2025;26(4):97-106.
8. Robert CP, Casella G. *Monte Carlo Statistical Methods*, 2nd Ed, Springer, 2004.
9. Pas VDS, Kleijn B, Vaart VDA. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron J Stat*. 2014;8(2):2585-2618.
10. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. 2017;27(5):1413-1432.
11. Zhang CH, Huang J. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann Stat*. 2008;36(4):1567-1594.