國立臺灣大學理學院統計與數據科學研究所

碩士論文

Institute of Statistics and Data Science

College of Science

National Taiwan University

Master's Thesis

高維度下 Spike-and-Slab 與 Horseshoe 先驗的穩健性比較

A Comparative Study of Spike-and-Slab and Horseshoe Prior for Robustness in High Dimension

練政楷

Cheng-Kai Lien

指導教授: 楊鈞澔 博士

Advisor: Chun-Hao Yang, Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202402347

Acknowledgements

在臺大統研所的這段學習旅程中，我經歷了許多困難和挑戰。這些困難不僅來自於學業壓力，還包括在論文研究過程中遇到的各種問題。然而，正是這些挑戰讓我更加堅定了自己的信念，並最終完成了這篇論文。這段時間對我來說是一段深刻的成長體驗，在此，我想藉此機會感謝在這過程中給予我幫助和支持的每一位。

首先，我要感謝我的父母，在這兩年中讓我可以專心完成碩士學業，無需為生活開銷而煩惱。你們的支持和栽培讓我能夠毫無後顧之憂地投入學習和研究。再來，我要感謝我的指導教授楊鈞澔 教授。這兩年來，您不厭其煩地回答我各種問題，並在我研究遇到瓶頸時，給予我方向和幫助。您的耐心指導和寶貴意見，是我在論文研究中克服困難、獲得進展的關鍵。每一次的討論和交流都使我受益匪淺。此外，我還要感謝統研所的各位老師，這兩年來悉心的教導讓我學到了許多寶貴的知識，並提升了我的學術水平。最後，我要感謝這兩年來統計所的同學們，大家在遇到困難時總是團結合作，互相幫助，共同度過了許多挑戰並獲得了成長。

我在此向所有在這些日子裡給予我幫助和支持的人致以最誠摯的謝意。

練政楷　謹誌

國立臺灣大學，中華民國一一三

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

摘要

在統計分析中，當觀測數量 (n) 大於變量數 (p) 的情況下，一般採用頻率學派的方法進行變量選擇，這些方法在大多數狀況，模型的表現都還不錯。然而，當數據為稀疏高維度時，即變量數 (p) 遠大於觀測數量 (n)，使用一般常用的傳統頻率學派的方法可能面臨一些挑戰。本文使用貝葉斯方法作為變量選擇的一種替代方案。貝葉斯方法通過引入先驗分佈，為處理稀疏高維度數據提供了一種靈活的做法。本文首先比較兩種貝葉斯先驗分佈，分別是 Spike-and-Slab 先驗和 Horseshoe 先驗。這兩種先驗分佈在處理稀疏高維度數據時各有優缺點。本文將探討不同情況下這兩種先驗分佈的表現差異。此外，在本研究中，我們將挑選較為穩健的先驗分佈，並探討使用「後驗極端值調整平均數」取代「後驗平均數」比較兩者在變量選擇上的差異。所謂後驗極端值調整平均數，是指我們採用公認的穩健方法來調整先前選定的穩健先驗分佈，以進行有依據的極端值調整操作。文章將分析在不同數據集及參數設定下，各方法的表現如何，並評估哪種方法對數據異常值的穩健性質更佳。本研究的目的為在稀疏高維度數據的變量選擇提供更有效且穩健的貝葉斯方法，並期望應用在實際資料時，提供實用的參考依據。

關鍵字：貝葉斯變量選擇、Spike-and-Slab 先驗、Horseshoe 先驗、穩健性質、高維數據

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

Abstract

When the number of observations (n) exceeds the number of variables (p), classic frequentist approaches for variable selection are commonly utilized, and they work well in the majority of situations. Traditional frequentist approaches, on the other hand, could be challenged when the data is sparse and high-dimensional, which means that the number of variables (p) greatly exceeds the number of observations (n). This research employs Bayesian approaches as an alternative way of variable selection. By incorporating prior distributions, Bayesian approaches provide a flexible method to dealing with sparse, high-dimensional data. This study begins by comparing two Bayesian priors: the Spike-and-Slab prior and the Horseshoe prior. Each of these priors has pros and cons when working with sparse, high-dimensional data. The study will investigate the performance differences between these two priors under a variety of situations. Furthermore, in this study, we will choose a more robust prior distribution and use the ”posterior winsorized mean” instead of the ”posterior mean,” comparing the differences in variable selection between

doi:10.6342/NTU202402347

the two methods. The ”posterior winsorized mean” refers to utilizing a recognized robust method to adjust the previously selected robust prior distribution, hence executing a justified winsorization rather than an arbitrary adjustment. This paper will compare the performance of various algorithms across different datasets and parameter settings, determining which method provides greater resistance against outliers. This study aims to provide a more effective and robust Bayesian approach for variable selection in sparse high-dimensional data, as well as practical advices for actual data applications.

Keywords: Bayesian variable selection, Spike-and-Slab prior, Horseshoe prior, robustness, high-dimension data

doi:10.6342/NTU202402347

Contents

Page

Acknowledgements **i**

摘要 **iii**

Abstract **v**

Contents **vii**

List of Figures **ix**

List of Tables **xi**

Chapter 1 Introduction **1**

1.1 Challenges in High-Dimensional Variable Selection 1

1.2 The Frequentist and Bayesian Approaches 2

1.3 Robust Regression Techniques for Outliers 3

1.4 Motivation and Purpose 4

Chapter 2 Methodology **7**

2.1 Spike-and-Slab Prior 7

2.2 Horseshoe Prior 9

2.3 Variable Selection Criteria 10

2.4 Bayesian Combined with Robust Methods 12

doi:10.6342/NTU202402347

Chapter 3 Simulation Study **17**

3.1 Compare the Spike-and-Slab & Horseshoe Prior 17

3.2 Comparison of the Performance of Spike-and-Slab Winsorized Means 23

Chapter 4 Real Data Applications **27**

4.1 Gene Expression - Obese Diabetic Mice 27

4.2 Gene Expression - NCI60 29

Chapter 5 Conclusion **31**

References **33**

doi:10.6342/NTU202402347

List of Figures

3.1 Comparison of TPR and MCC at Different Signal-to-Noise Ratios 19

3.2 Comparison of TPR and MCC at Different Sparsity Levels 21

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

List of Tables

2.1 Confusion Matrix 11

3.1 Confusion Matrix Comparison at Varying SNR 19
3.2 Confusion Matrix Comparison at Varying SP 20
3.3 Performance Metrics by Case and Contamination Ratio: SS vs H with Outliers 22
3.4 Performance Metrics by Case and Contamination Ratio: Comparing Spike-and-Slab Winsorized Mean with Outliers to Other Methods 24
3.5 Performance Metrics by Case and Contamination Ratio: Comparing Spike-and-Slab Winsorized Mean with High Leverage to Other Methods 26

4.1 Performance Metrics for ”Obese Diabetic Mice” Simulation 28
4.2 Selected Variables for Different Responses Using Various Methods-Obese Diabetic Mice 28
4.3 Performance Metrics for ”NCI60” Simulation 30
4.4 Selected Variables Across Different Methods-NCI60 30

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

Chapter 1 Introduction

1.1 Challenges in High-Dimensional Variable Selection

High-dimensional data analysis is becoming increasingly popular in data science, particularly in domains such as financial analysis, genomics, social network research, and machine learning. These datasets can contain a huge numbers of variables, challenging the construction and interpretation of statistical models (Fan and Li, 2006). Bellman (1957) was credited with coining the phrase ”curse of dimensionality,” which mainly refers to the possibility of an exponential rise in computing costs with an increase in variables (dimensions). This issue is most obvious in high-dimensional situations, when there are much less samples available than there are variables. This results in unstable model estimation and worse reliability. Additionally, in high-dimensional data analysis, variable selection plays a critical role in statistical inference. This section will explore numerous issues faced during the variable selection procedure in high-dimensional data.

Multiple variables with low levels of noise in data analysis might cause ”noise accumulation,” increasing their combined influence on the research. This problem becomes more acute with high-dimensional data, where identifying and analyzing relevant characteristics is difficult, especially when the signals of critical variables are weak. Fan and Fan (2008) emphasized the difficulties involved in high-dimensional statistical analysis.

doi:10.6342/NTU202402347

In many high-dimensional applications, only a few essential factors have an important impact on the result, whereas a huge number of irrelevant variables not only provide little benefit but also increase the noise and complexity of the study. Furthermore, according to Fan et al. (2014), employing sparse models can help relieve the challenges discussed above.

Fan and Lv (2010) and Fan et al. (2014) provided an overview of some of the issues and challenges related to high-dimensional data. The problems with high-dimensional data that were previously discussed could prevent typical model selection criteria from obtaining appropriate statistical inferences. Chen et al. (2014) have pointed out that in high-dimensional situations, criteria like AIC and BIC could select too many variables. As a result, we must employ additional statistical approaches (Section 1.2) to help with variable selection in high dimensions.

1.2 The Frequentist and Bayesian Approaches

To avoid overfitting and improve generalizability, frequentist researchers employ several regularization techniques for high-dimensional variable selection. For instance, a popular technique called Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani (1996)) selects variables and regularizes them by applying an absolute penalty to the coefficients. Other methods include Adaptive Lasso (Zou, 2006), which uses different penalty strengths for different coefficients, Smoothly Clipped Absolute Deviation (SCAD, Fan and Li (2001)) and Minimax Concave Penalty (MCP, Zhang (2010)), both of which use nonlinear penalties to avoid over-penalizing large coefficients. Furthermore, a method designed especially for variable selection in ultrahigh-dimensional data is Sure

doi:10.6342/NTU202402347

Independence Screening (SIS, Fan and Lv (2008)). SIS first filters out some unimportant variables, and then use the previously mentioned approaches to continue variable selection. In the frequentist approaches, each of these techniques is a useful tool for variable selection.

Bayesian variable selection procedures allow us to employ methods like Bayes factor (Kass and Raftery, 1995) and Bayesian Model Averaging (BMA, Raftery et al. (1997)) in high-dimensional data analysis, including shrinkage priors. Shrinkage priors such as the Laplace prior (Park and Casella, 2008), Spike-and-Slab prior (George and McCulloch, 1993), and Horseshoe prior (Carvalho et al., 2010), with a focus on the Spike-and-Slab prior and Horseshoe prior in Chapter 2. To decide whether a variable should be included in the model, MCMC methods can be used to estimate the posterior inclusion probability.

Additionally allowing the flexibility to select the model's priors, Bayesian variable selection has a number of advantages, including the capacity to estimate the posterior probability of every feasible model (O'Hara and Sillanpää, 2009). The predicted model can approach the true model even for high-dimensional sparse data by selecting appropriate variables, and its accuracy increases with the size of the data.

1.3 Robust Regression Techniques for Outliers

Dealing with outliers or leverage points can be challenging in data analysis. They might result from inaccurate measurements, incorrect data input, or real observations that deviate from the model's assumptions. It becomes more challenging in such situations to find the truly important variables. Robust variable selection methods have been presented as solutions to this problem. These techniques aim to make variable selection more robust

doi:10.6342/NTU202402347

with contaminated data.

Serneels et al. (2005) devised a method known as ”Partial Robust M-regression” (PRM) to improve the robustness of Partial Least Squares (PLS, Wold (1985)) regression against outliers. The PRM technique adds a ”weighting” mechanism to the PLS framework, considering both independent and dependent variables. By assigning lower weights to outliers, their impact on the model can be effectively reduced; robust LASSO regression (Alfons et al., 2013) combines the characteristics of sparse Least Trimmed Squares (LTS, Rousseeuw (1984)), considering only the lowest squared residuals, allowing for the elimination of some outliers. The core concept of robust elastic net regression (Kurnaz et al., 2018) is similar to that of robust LASSO regression, also incorporating sparse LTS characteristics, but differs in the penalty used. Overall, these methods primarily involve reducing the weight of outliers or trimming some outliers. Both aim to reduce the influence of outliers.

Filzmoser and Nordhausen (2021) gave a robust high-dimensional overview, noting that there are significantly more academic papers on how to maintain ”robustness” in models when $n > p$ than when $p > n$. This is due to the fact that variable selection is more difficult when dealing with high-dimensional data containing outliers. Filzmoser and Nordhausen (2021) also pointed out that variable selection approaches in high-dimensional situations are often useful when $n > p$.

1.4 Motivation and Purpose

In high-dimensional Bayesian variable selection, both Spike-and-Slab and Horseshoe priors play important roles and are widely considered to be particularly effective for this

doi:10.6342/NTU202402347

purpose. The Spike-and-Slab prior, with its unique structure, efficiently distinguishes between significant and insignificant variables, making it an ideal choice for dealing with sparsity issues in high-dimensional data. This prior categorizes variables into two groups: "slab", which denotes variables with significant influence on the model, and "spike" which refers to variables with minimal influence. This obvious distinction helps in more accurately identifying important variables.

On the other hand, the Horseshoe prior produces sparsity via a continuous shrinking effect. For significant variables, the Horseshoe prior gives enough flexibility to keep them away from zero, allowing them to be properly identified as signals. Meanwhile, it imposes a large shrinkage effect on the insignificant variables, driving them towards zero and help in achieving model sparsity.

The purpose of this study is to compare these two priors in Bayesian variable selection, particularly in high-dimensional scenarios where the data is contaminated. This comparison aims to determine which prior performs better and to enhance the robustness of the more effective approach. The primary reason for comparing the Spike-and-Slab and Horseshoe priors is their different strategies for identifying key variables: the Spike-and-Slab employs a straightforward classification method to identify significant variables, while the Horseshoe utilizes its continuous shrinkage properties. By comparing their performance across various scenarios, we can ascertain which prior is more suitable to specific applications, thus improving the accuracy of variable selection and the usefulness of the model, and select one of the priors to make it more robust.

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

Chapter 2 Methodology

In Section 2.1, we introduced the Spike-and-Slab prior, as well as the method proposed by Biswas et al. (2022) to save computational time. In Section 2.2, we introduced the Horseshoe prior and appropriate hyperparameters of the prior selection. In Section 2.3 we introduced criteria for variable selection after the parameter estimation. In Section 2.4, we introduced the Robust Least Angle Regression (RLARS, Khan et al. (2007)) method and combined it with Bayesian approaches to improve the robustness.

2.1 Spike-and-Slab Prior

We primarily consider a standard linear regression model for data $Y \in \mathbb{R}^n$ as the response variable. Furthermore, we concentrate on high-dimensional data, that is, $n \ll p$. Our likelihood and prior are the following

$$Y|X, \beta, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(X\beta, \sigma^2 I_n) \tag{2.1}$$

$$\beta_j|\gamma_j, \sigma^2 \overset{\text{ind}}{\sim} (1-\gamma_j)\mathcal{N}(0, \sigma^2{\tau_0}^2) + \gamma_j\mathcal{N}(0, \sigma^2{\tau_1}^2), \quad j = 1, \ldots, p \tag{2.2}$$

$$\gamma_j \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(q), \quad j = 1, \ldots, p \tag{2.3}$$

$$\sigma^2 \sim \text{InvGamma}\left(\frac{\alpha_0}{2}, \frac{\alpha_1}{2}\right) \tag{2.4}$$

doi:10.6342/NTU202402347

where $X \in \mathbb{R}^{n \times p}$ is the design matrix, $q \in (0,1)$, ${\tau_1}^2 \gg {\tau_0}^2$ and $\alpha_0, \alpha_1 > 0$. It indicates that the jth covariate has been included to the model if $\gamma_j = 1$. In contrast, the jth covariate should be removed from the model if $\gamma_j = 0$. The "spike" and "slab" components of the prior are represented by $\mathcal{N}(0, \sigma^2\tau_0^2)$ and $\mathcal{N}(0, \sigma^2\tau_1^2)$, respectively.

Spike-and-Slab prior is an essential technique in the Bayesian tools for variable selection, valued for their interpretability and decisive posterior probability analysis. The method was derived from George & McCulloch's seminal work (1993). This method is commended for making it simple to measure each predictor's contribution inside a model, which makes the model selection process easier to understand (George and McCulloch, 1993, 1997). However, the method is not without challenges; computing demands may arise during implementation. The Spike-and-Slab approach can be computationally intensive, especially when dealing with large datasets, as Bai et al. (2021) explained. This makes it unfeasible for complex statistical analyses.

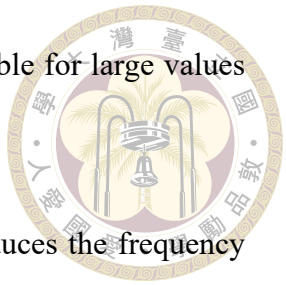
Next, we will introduce the "Scalable Spike-and-Slab" (S^3) method proposed by Biswas et al. (2022), which saves computational time for Spike-and-Slab. Sampling β ($\beta \in \mathbb{R}^p$) from the full conditional distribution $\pi(\beta|\gamma, \sigma^2, y)$ is the primary source of computational difficulties in Gibbs samplers. The full conditional distribution is a multivariate normal distribution,

$$\beta|\gamma, \sigma^2, y \sim \mathcal{N}(\Sigma^{-1}X^TY, \sigma^2\Sigma^{-1}) \tag{2.5}$$

where $\Sigma = X^TX + D$, D is the diagonal matrix, and the vector $\gamma\tau_1^{-2} + (1_p - \gamma)\tau_0^{-2}$ populates its diagonal members. Direct sampling from this distribution involves computing the inverse of Σ, which incurs a computational cost of $\Omega(p^3)$ and becomes prohibitive with large p. By implementing the method proposed by Bhattacharya et al. (2016), the

doi:10.6342/NTU202402347

complexity can be reduced to $\Omega(n^2 p)$, although it remains considerable for large values of p.



S^3 introduces a pre-computation approach that significantly reduces the frequency and complexity of recalculations needed in each Gibbs sampling iteration, particularly for the matrix Σ and its inverse Σ^{-1}. This change simplifies sampling from the distribution of β, lowering total computational costs.

By utilizing pre-computation and reuse the state from the previous iteration, S^3 avoids the need to recalculate the pre-computed matrix from the beginning in each iteration, which significantly reduces the complexity. This computational method is especially advantageous in high-dimensional data settings where p is large. For a linear regression model, S^3 greatly minimizes computing complexity while maintaining the quality of the sample by carefully using pre-computation and the output from the preceding iteration.

2.2 Horseshoe Prior

The Horseshoe prior (Carvalho et al., 2010) captures sparse signals well by implementing major shrinkage on less relevant parameters and providing considerable flexibility for the important ones. A standard linear regression model is also taken into consideration for the response variable, $Y \in \mathbb{R}^n$. Additionally, we focus on high-dimensional data, or $n \ll p$. Below are the prior and the likelihood.

$$Y|X, \beta, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(X\beta, \sigma^2 I_n) \tag{2.6}$$

$$\beta_j \mid \lambda_j^2, \tau^2, \sigma^2 \sim N(0, \lambda_j^2 \tau^2 \sigma^2), \quad j = 1, \ldots, p \tag{2.7}$$

$$\lambda_j \overset{\text{iid}}{\sim} C^+(0, 1), \quad j = 1, \ldots, p \tag{2.8}$$

doi:10.6342/NTU202402347

$$\tau \sim C^{+}(0, \kappa) \tag{2.9}$$

$$\sigma^2 \sim 1/\sigma^2 \tag{2.10}$$

where the termed τ is global shrinkage parameter, and λ_j is called the local shrinkage parameter, which allows for distinct shrinkage coefficients for different β_j. All weights are reduced to zero by the global parameter τ, but some weights are allowed to escape because the local scales λ_j contain heavy half-cauchy tails. The default value for κ is 1, yet this is generally not the best option. Piironen and Vehtari (2017) provided a procedure for choosing κ. Let

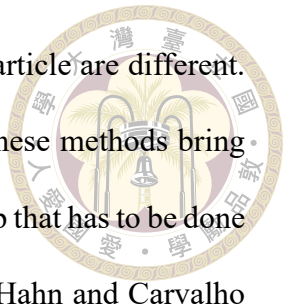
$$T = \frac{S}{p - S}\frac{\sigma}{\sqrt{n}}, \quad \kappa = \frac{T^2}{\sigma^2} \tag{2.11}$$

where p is the number of variables and S is the number of non-zero β entires, that is, significant coefficients. In practice, the exact value of S is unknown. As a result, S is estimated based on our understanding of the data, reflecting our prior assumptions about the model's complexity, namely sparsity. This method includes prior knowledge of the sparsity of the model into the model itself. Even if the estimation of S is very rough, choosing the prior for τ in this way, where κ also takes into account the sample size (n) and noise (σ), is better than simply assuming τ follows a $C^{+}(0, 1)$ distribution.

2.3 Variable Selection Criteria

In Chapter 1, approaches like LASSO and SCAD are mentioned. These methods set the estimates of insignificant variables to zero after completing parameter estimation. Therefore, variables that have a non-zero estimated value are considered significant. In this sense, these methods achieve variable selection in addition to parameter estimation.

doi:10.6342/NTU202402347



However, the Spike-and-Slab prior and Horseshoe prior used in this article are different. Instead of setting the estimates of insignificant parameters to zero, these methods bring them near to zero. This means that variable selection is a necessary step that has to be done in Sections 2.1 and 2.2 once parameter estimation is finished. Both Hahn and Carvalho (2015) and Ghosh et al. (2019) have proposed methods for variable selection. After taking these methods into consideration, we decide to select variables using the Signal Adaptive Variable Selector (SAVS) technique suggested by Ray and Bhattacharya (2018) (see Algorithm 1). This method is not only easy to use but also operates quickly and produces satisfactory results. The underlying idea is that this approach permits relatively insignificant variables to be reduced to zero after getting the β posterior sample using a continuous shrinkage prior. Therefore, the non-zero $\hat{\beta}^*$ are considered important, which successfully divides the variables into signal and noise.

Algorithm 1: SAVS Algorithm

Require: Posterior mean $\hat{\beta}$ and design matrix X

for $j = 1$ to p **do**
  $\mu_j = 1/|\hat{\beta}_j|^2$
  if $|\hat{\beta}_j| \cdot \|X_j\|^2 \leq \mu_j$ **then**
    $\hat{\beta}_j^* = 0$
  else
    $\hat{\beta}_j^* = \text{sign}(\hat{\beta}_j)\|X_j\|^{-2}\left(|\hat{\beta}_j| \cdot \|X_j\|^2 - \mu_j\right)$
  end if
end for

Ensure: A sparse estimate $\hat{\beta}^*$

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 2.1: Confusion Matrix

The next step is to assess and compare these two priors using various metrics after variable selection using the SAVS method. Confusion matrix is displayed in Table 2.1. We concentrate on three key performance indicators: the true positive rate (TPR), the false

doi:10.6342/NTU202402347

discovery rate (FDR) and Matthew's correlation coefficient (MCC). The true positive rate ranges from 0 to 1, where TPR = 1 indicates the full accuracy of all true positive case predictions and TPR = 0 indicates no true positive case predictions at all. The false discovery rate ranges from 0 to 1, where FDR = 1 indicates that all predicted positive cases are actually false positives, and FDR = 0 indicates that none of the predicted positive cases are false positives. The Matthew's correlation coefficient ranges from -1 to 1, where a value of 1 indicates excellent predictions, 0 indicates random predictions, and -1 indicates entirely wrong predictions. While MCC is concerned with the overall prediction outcomes, TPR is more concerned with the capacity to detect positive cases. TPR is relatively more crucial than the others as it is our goal to choose as many significant variables from the actual data as possible. The TPR is defined by TP/(TP + FN), the FDR is defined by FP/(TP + FP) and MCC is defined as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.12}$$

Our objective is to compare the Spike-and-Slab prior with the Horseshoe prior in terms of TPR, FDR and MCC under different settings, as well as to assess their robustness to contaminated data. To improve the robustness of our analysis conclusions, we suggest adopting the posterior winsorized mean (Wilcox, 2005) (see Section 2.4) rather than the standard posterior mean.

2.4 Bayesian Combined with Robust Methods

Least angle regression (LARS, Efron et al. (2004)) is a linear regression model selection algorithm. It develops the model sequentially from scratch, selecting the variable

doi:10.6342/NTU202402347

with the highest correlation with the current residuals at each stage and adjusting in the same direction until another variable has an equivalent level of correlation with the residuals. The LARS algorithm is especially effective at dealing with high-dimensional data and provides a clear ranking of variable importance.

Robust LARS (RLARS, Khan et al. (2007)) is an extension to the LARS method designed to improve model robustness. This enhancement incorporates robust correlation and winsorization techniques (Wilcox, 2005). The process is as follows: First, RLARS calculates a robust correlation matrix through adjusted winsorization. Then, it defines a tolerance ellipse using the robust correlation matrix and applies winsorization to data points that fall outside the ellipse, causing them to move to the ellipse's boundary. Next, RLARS computes robust correlation using winsorized data points, which replaces the Pearson correlation originally used in LARS. This adjustment not only reduces the impact of outliers on variable selection but also allows RLARS to effectively derive a "reduced set" of important variables through its selection process. Finally, robust parameter estimation approaches, such as MM-estimate (Yohai, 1987) or Least Trimmed Squares (LTS, Rousseeuw (1984)) estimators, are used to estimate the coefficients of the reduced set.

Based on the results from Table 3.3 comparing Spike-and-Slab with Horseshoe, we choose to combine the Spike-and-Slab prior with RLARS. Detailed discussion of Table 3.3 will be provided in Chapter 3. The motivation for considering the combination of the two methods is to compensate for each other's weaknesses. The RLARS approach is known for its robustness and great computational efficiency, making it ideal for dealing with high-dimensional data. However, even in simple situations, this method can select too many variables for the model. On the other hand, as mentioned in Section 2.1, the Scalable Spike-and-Slab (S^3) enhances the computational efficiency of the Spike-and-Slab method

doi:10.6342/NTU202402347

in high-dimensional data. Additionally, being a Bayesian approach, Sike-and-Slab provides a complete posterior distribution of information. However, its main drawback is that it is not robust enough, making it susceptible to outliers.

Algorithm 2: Winsorization Algorithmn

Require: $\boldsymbol{\beta}$ is a vector of length m, lower percentile L, upper percentile U

1: For each $\beta^{(i)}$ as follows:

$$\beta_{\mathbf{W}}^{(i)} = \begin{cases} \beta_L & \text{if } \beta^{(i)} < \beta_L \\ \beta^{(i)} & \text{if } \beta_L \leq \beta^{(i)} \leq \beta_U \\ \beta_U & \text{if } \beta^{(i)} > \beta_U \end{cases} \tag{2.13}$$

for $i = 1, \ldots, m$, where $\beta_L = \text{quantile}(\{\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(m)}\}, L)$, $\beta_U = \text{quantile}(\{\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(m)}\}, 1 - U)$

2: Compute the average:$\hat{\beta}_{\mathbf{w}} = \frac{1}{m} \sum_{i=1}^{m} \beta_{\mathbf{w}}^{(i)}$

Ensure: Winsorized averaged value $\hat{\beta}_{\mathbf{w}}$

Rather than simply combining Spike-and-Slab with RLARS, it would be more accurate to say that the RLARS method is used to enhance the robustness of the Spike-and-Slab approach. For detailed steps, see Algorithm 3.

The main concept behind "Spike-and-Slab Winsorized Means" is that after completing MCMC, the Spike-and-Slab approach produces a posterior sample for each variable after m iterations. These variables are then subjected to various winsorizing treatments before averaging, with the aim of increasing the values of important variables and decreasing the values of less important variables. The RLARS method is used to determine which variables are important and which are not. Additionally, since RLARS tends to select more variables, the coefficients estimated are classified into two groups using k-means. By using k-means to split the coefficients into two cluster, the method identifies sparse important variables in high-dimensional data. Clusters with fewer elements are consid-

doi:10.6342/NTU202402347

Algorithm 3: Spike-and-Slab Winsorized means

Require:
Posterior samples $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p\}$ with m iterations, posterior mean $\hat{\beta}_{SS}$, RLARS estimates $\hat{\beta}_{\text{RLARS}}$

Compute $N_{\text{RLARS}} = \#\left\{\hat{\beta}_{\text{RLARS}} \neq 0\right\}$
Apply K-means clustering on $\hat{\beta}_{\text{RLARS}}$ to form two groups
$N_{K-\text{RLARS}}$ = size of smaller group from K-means
Determine the set of final variables S_{final} of size $N_{\text{final}} = \min\left(N_{\text{RLARS}}, N_{K-\text{RLARS}}\right)$
for $v = 1, 2, \ldots, p$ **do**
  if $v \in S_{\text{final}}$ **then**
    if $\hat{\beta}_{SS}(v) > 0$ **then**
      Winsorization lower percentile to $L = 0.6$ and upper percentile to $U = 0$
    else
      Winsorization lower percentile to $L = 0$ and upper percentile to $U = 0.6$
    end if
  else
    if $\hat{\beta}_{SS}(v) > 0$ **then**
      Winsorization lower percentile to $L = 0$ and upper percentile to $U = 0.6$
    else
      Winsorization lower percentile to $L = 0.6$ and upper percentile to $U = 0$
    end if
  end if
end for
Compute the Winsorized posterior mean using Algorithm 2 to get $\hat{\beta}_{\text{w}}$
Apply SAVS (algorithm 1) to $\hat{\beta}_{\text{w}}$ to obtain sparse estimates $\hat{\beta}^*$
Select the final variables with non-zero $\hat{\beta}^*$ entires

Ensure: The final set of selected variables

ered more significant and are compared with the initial number of non-zero coefficients estimated by RLARS. The set with the fewer elements is deemed the important variables utilized to adjust the posterior samples.

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

Chapter 3 Simulation Study

In Section 3.1, we compare the Spike-and-Slab prior and the Horseshoe prior in different scenarios. In Section 3.2, we compare the performance of Spike-and-Slab Winsorized means ($\mathbf{SS_W}$) with other methods under various conditions.

3.1 Compare the Spike-and-Slab & Horseshoe Prior

The performance of the Spike-and-Slab prior and the Horseshoe prior in different situations will be examined and compared. We are able to comprehend these two prior distributions' applicability as well as their pros and cons under various statistical models and datasets through the of these comparisons. Next, we will investigate how well the Spike-and-Slab prior and Horseshoe prior perform at various signal-to-noise ratio (SNR) and sparsity levels (sp). The definitions of SNR and SP are as follows:

$$\text{SNR} = \frac{\text{tr}(\text{Var}(\boldsymbol{X}\boldsymbol{\beta}))}{\sigma^2}, \quad \text{SP} = \frac{S}{p} \tag{3.1}$$

where S denotes the number of non-zero $\boldsymbol{\beta}$ coefficients, and p refers to the total number of features, that is, the length of $\boldsymbol{\beta}$.

We evaluate the performance of Spike-and-Slab prior and Horseshoe prior at various

doi:10.6342/NTU202402347

SNR. In each scenario, we will run 100 simulations. The following Spike-and-Slab prior settings are all based on the parameters listed below:

- $\alpha_0 = \alpha_1 = {\tau_1}^2 = 1, \quad {\tau_0}^2 = \frac{1}{\sqrt{n}}$ (corresponding to equation (2.2),(2.4))
- Defined as $K = \max(10, \log(n))$.
- Compute a sequence of q values ranging from 0.0001 to (1 - 0.0001) with a step size of 0.0001. (q corresponding to equation (2.3))

$$q = \begin{cases} \frac{1}{p} & \text{if multiple } q \text{ values minimize } |F(K; p, q) - 0.9|, \\ q_{\text{min}} & \text{otherwise,} \end{cases}$$

where q_{min} is the value in the sequence that minimizes $|F(K; p, q) - 0.9|$, and $F(K; p, q)$ denotes the binomial probability of K given p and q. Here, $F(\cdot)$ represents the cumulative distribution function (CDF). Additionally, the following Horseshoe prior settings are based on the equations (2.6) to (2.11). After calculating the posterior mean for both Spike-and-Slab prior and Horseshoe prior, we use SAVS (Algorithm 1) for variable selection.

In Table 3.1 and Figure 3.1, we set $n = 200$, $p = 500$, $S = 25$ (25 of the β_{true} are equal to 1, and the remaining $p - S$ are 0.) Repeat the simulation 100 times. We can observe from Table 3.1 and Figure 3.1 that both priors' performance deteriorate with a drop in SNR. It is clear from Figure 3.1 that both priors show strong TPR when the SNR is quite high ($\sigma = 1$). When $\sigma = 2$, we can see that both methods have a TPR close to 1, but the Spike-and-Slab has a lower FDR. When $\sigma = 3$, the Horseshoe shows a higher TPR, but correspondingly, it also has a higher FDR. When $\sigma = 4$, the TPRs of both methods are similar, but the Spike-and-Slab has a slightly lower FDR. From the right plot, we can observe that except for $\sigma = 1$, where both methods have an MCC of 1, the Spike-and-Slab

doi:10.6342/NTU202402347

σ	SNR	Spike-and-Slab		Horseshoe	
1	25.02	TP: 25 (0) FP: 0.02(0.14) FDR: **0** MCC: **1**	FN: 0 (0) TN: 474.98 (0.14)	TP: 25 (0) FP: 1 (0.97) FDR: 0.04 MCC: 0.98	FN: 0 (0) TN: 474 (0.97)
2	6.23	TP: 23.86 (2.1) FP: 2.18 (2.74) FDR: **0.08** MCC: **0.93**	FN: 1.14 (2.1) TN: 472.82 (2.74)	TP: 24.89 (0.35) FP: 21.66(4.79) FDR: 0.46 MCC: 0.72	FN: 0.11 (0.35) TN: 453.34(4.79)
3	2.79	TP: 19.8 (2.33) FP: 19.73 (6.69) FDR: **0.49** MCC: **0.61**	FN: 5.2 (2.33) TN: 455.27(6.69)	TP: 22.53 (1.64) FP: 53.14 (7.93) FDR: 0.7 MCC: 0.48	FN: 2.47(1.64) TN: 421.86(7.93)
4	1.56	TP: 19.01(2.39) FP: 44.37 (9.32) FDR: **0.7** MCC: **0.44**	FN: 5.99 (2.39) TN: 430.63(9.32)	TP: 19.7 (2.14) FP: 81.08(9.55) FDR: 0.8 MCC: 0.34	FN: 5.3 (2.14) TN: 393.92 (9.55)

Table 3.1: The performance of the Spike-and-Slab prior and the Horseshoe prior using confusion matrices at different SNR. The standard deviations are shown in parentheses, with σ values ranging from 1 to 4.

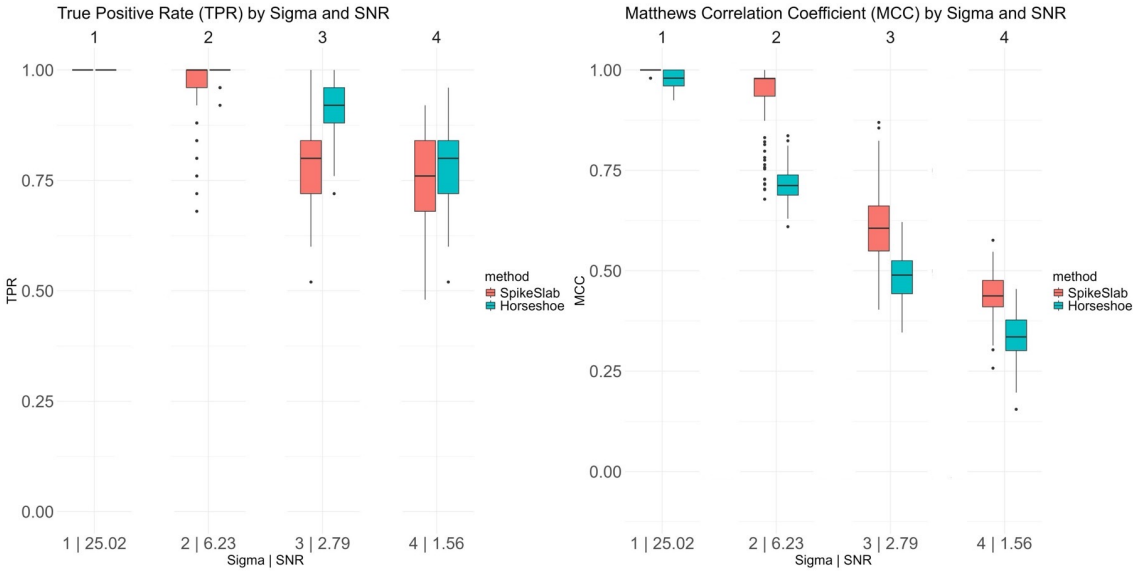


Figure 3.1: The left graph shows box plots of the TPR for Spike-and-Slab prior as well as Horseshoe prior at various SNR based on 100 simulations. The right graph displays box plots of MCC under the same SNR settings, which is similarly based on 100 simulations. The red box represents the Spike-and-Slab prior, while the blue box represents the Horseshoe prior.

generally has a higher MCC than the Horseshoe in other situations. Overall, although the Spike-and-Slab method performs slightly better than the Horseshoe method, the difference between them is not significant. As a result, there is no need to consider SNR size while

doi:10.6342/NTU202402347

determining which approach should be used to practical applications.

S	SP	Spike-and-Slab		Horseshoe	
10	0.02	TP: 10 (0) FP: 0.07 (0.26) FDR: **0.01** MCC: **1**	FN: 0 (0) TN: 489.93(0.26)	TP: 10 (0) FP: 0.24 (0.51) FDR: 0.02 MCC: 0.99	FN: 0 (0) TN: 489.76(0.51)
50	0.1	TP: 37.26 (3.61) FP: 44.48 (9.1) FDR: 0.54 MCC: 0.53	FN: 12.74 (3.61) TN: 405.52(9.1)	TP: 43.56 (3.56) FP: 47.88(8.92) FDR: **0.52** MCC: **0.6**	FN: 6.44(3.56) TN: 402.12 (8.92)
100	0.2	TP: 68.31 (5.24) FP: 106.98 (12.3) FDR: **0.61** MCC: **0.35**	FN: 31.69 (5.24) TN: 293.02 (12.3)	TP: 60.78(5.85) FP: 94.05(10.88) FDR: **0.61** MCC: 0.32	FN: 39.22(5.85) TN: 305.95(10.88)
150	0.3	TP: 101.4(6.96) FP: 130.39 (11.32) FDR: **0.56** MCC: **0.28**	FN: 48.6 (6.96) TN: 219.61(11.32)	TP: 85.76 (8.12) FP: 108.78 (11.59) FDR: **0.56** MCC: 0.25	FN: 64.24 (8.12) TN: 241.22 (11.59)

Table 3.2: The performance of the Spike-and-Slab prior and the Horseshoe prior using confusion matrices at different sparsity (SP). The standard deviations are shown in parentheses, with SP values ranging from 0.02 to 0.3.

In Table 3.2 and Figure 3.2, we set $n = 200, p = 500$, S represents the number of non-zero β_{true} values, which are equal to 1, and the remaining $p-S$ are 0. Repeat the simulation 100 times. Based on the previous results, we know that as the SNR decreases, the quality of variable selection also deteriorates. Therefore, in the four scenarios mentioned above, we fixed the SNR at 10 and focused solely on the performance of both priors as sparsity increases. We can observe from Table 3.2 and Figure 3.2 that both priors' performance deteriorates with an increase in sparsity value (SP). It is clear from Figure 3.2 that both priors show strong TPR and MCC when the SP is quite low (sp = 0.02). When sp = 0.1, the Horseshoe prior performs better than the Spike-and-Slab prior, with a slightly lower FDR, slightly higher TPR, and MCC. When sp = 0.2 and 0.3, the FDR of the Spike-and-Slab and Horseshoe priors are similar, but the Spike-and-Slab has slightly higher TPR and MCC. Overall, the differences between the two priors are insignificant. In practical applications,

doi:10.6342/NTU202402347

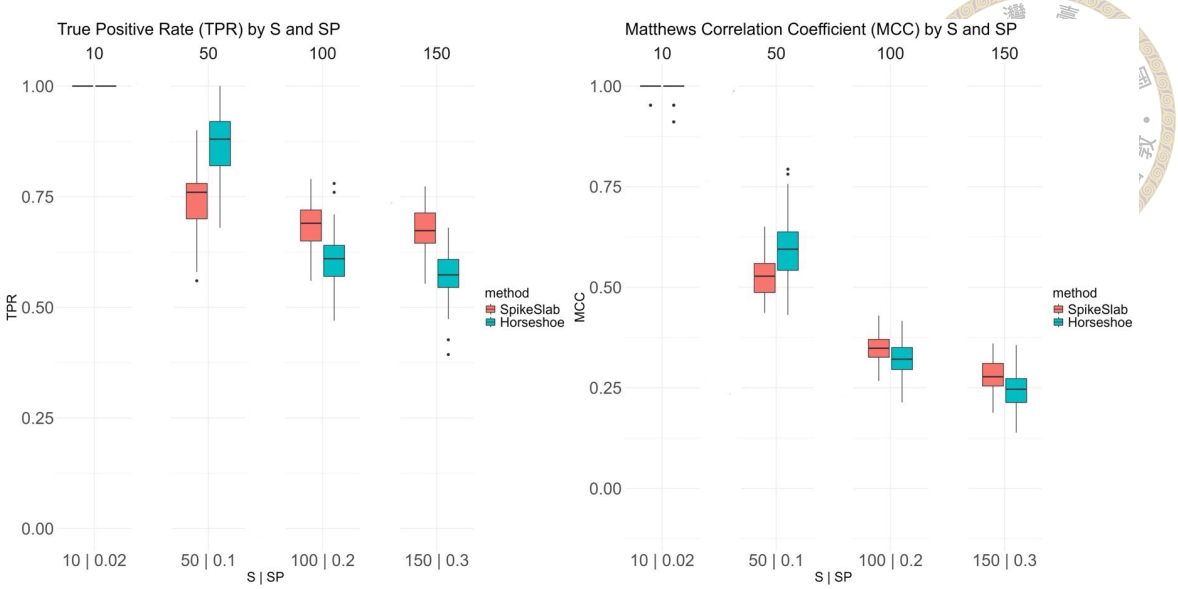


Figure 3.2: The left graph shows box plots of the TPR for Spike-and-Slab prior as well as Horseshoe prior at various sparsity level (SP) based on 100 simulations. The right graph displays box plots of the MCC under the same sparsity settings, which is similarly based on 100 simulations."S" represents the number of non-zero β. The red box represents the Spike-and-Slab prior, while the blue box represents the Horseshoe prior.

there is no need to decide an approach based on sparsity level.

Comparison for Robustness

Next, we will compare the performance of Spike-and-Slab and Horseshoe when dealing with outliers. We simulate outliers by contaminating the data under different conditions. As before, we first calculate the posterior means of Spike-and-Slab and Horseshoe and then apply SAVS (Algorithm 1). The SAVS used here is set to its default settings. We set $n = 200$, $p = 500$ and the number of non-zero β_{true} values to 25, with 15 values equal to 1 and 10 values equal to -1. This simulates the presence of both positively and negatively correlated signals. Similarly, we simulate linear regression as follows:

$$Y = X\beta + \epsilon \tag{3.2}$$

We consider different distribution scenarios to contaminate the data and simulate outliers:

doi:10.6342/NTU202402347

- A. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\text{Cauchy}(0,1)$, heavy-tailed Cauchy contamination;
- B. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\mathcal{N}(5,1)$, shifted Normal contamination;
- C. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\mathcal{N}(0,5)$, high variance Normal contamination;

where α is the proportion of contaminated data.

case:		**A**		**B**		**C**	
contamination		**SS**	**H**	**SS**	**H**	**SS**	**H**
	TPR:	0.93	0.96	1	1	1	1
5%	MCC:	**0.82**	0.8	**0.99**	0.86	**0.99**	0.87
	FDR:	0.2	0.28	0.01	0.24	0.01	0.22
	TPR:	0.91	0.9	0.97	0.99	0.96	0.99
10%	MCC:	**0.79**	0.59	**0.96**	0.74	**0.94**	0.77
	FDR:	0.24	0.52	0.05	0.43	0.08	0.38
	TPR:	0.83	0.83	0.83	0.97	0.85	0.97
20%	MCC:	**0.52**	0.41	**0.75**	0.6	**0.78**	0.61
	FDR:	0.51	0.69	0.29	0.59	0.26	0.57

Table 3.3: Comparing Spike-and-Slab and Horseshoe performance under different contamination scenarios (5%, 10%, and 20%). Each situation is simulated 100 times, with the results averaged.

The results in Table 3.3 show that the Horseshoe prior consistently has a higher FDR than the Spike-and-Slab prior in all three cases. When only 5% contamination, both approaches have a high TPR, but the Spike-and-Slab prior has a lower FDR. When contamination reaches 20%, the TPR of the Horseshoe prior is sometimes higher, but its FDR is also relatively higher. Therefore, the MCC of the Horseshoe prior is lower than that of the Spike-and-Slab prior. Overall, first, the Spike-and-Slab method outperforms the Horseshoe method. Second, neither method can be considered robust, but if we must compare, the Spike-and-Slab method is relatively more robust. Third, the scalable Spike-and-Slab is 5-10 times faster than the Horseshoe method, and computational speed is especially important in high-dimensional settings. Based on these three reasons, we choose the Spike-and-Slab method for further improvement to enhance its robustness.

doi:10.6342/NTU202402347

3.2 Comparison of the Performance of Spike-and-Slab Winsorized Means

In Section 3.1, we selected the Spike-and-Slab prior distribution to be improved and enhanced its robustness using the RLARS algorithm. For the detailed approximation process, please refer to Algorithm 3 (Spike-and-Slab Winsorized Means). Next, we will observe the performance of the Spike-and-Slab Winsorized Means. We also set $n = 200$, $p = 500$ and the number of non-zero β_{true} values to 25, with 15 values equal to 1 and 10 values equal to -1. Similarly, we simulate linear regression as follows:

$$Y = X\beta + \epsilon \tag{3.3}$$

Comparison for Robustness Against Outliers

We consider different distribution scenarios to contaminate the data and simulate outliers:

- A. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\text{Cauchy}(0,1)$, heavy-tailed Cauchy contamination;
- B. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\mathcal{N}(5,1)$, shifted Normal contamination;
- C. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\mathcal{N}(0,5)$, high variance Normal contamination;
- D. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\text{Laplace}(0,8)$, heavy-tailed Laplace contamination;
- E. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\text{Lognormal}(0,2)$, heavy-tailed log-normal contamination.

where α is the proportion of contaminated data.

doi:10.6342/NTU202402347

case:		A				B				C			
contamination		SS_W	SS	RLARS	MCP	SS_W	SS	RLARS	MCP	SS_W	SS	RLARS	MCP
	TPR:	0.99	0.93	1	0.95	1	1	1	1	1	1	1	1
5%	MCC:	**0.96**	0.82	0.76	0.57	**1**	0.99	0.74	0.58	**1**	0.99	0.75	0.58
	FDR:	0.04	0.2	0.39	0.62	0	0.01	0.42	0.63	0	0.01	0.4	0.63
	TPR:	0.98	0.91	1	0.81	0.99	0.97	0.99	1	0.99	0.96	1	1
10%	MCC:	**0.85**	0.79	0.77	0.43	**0.98**	0.96	0.74	0.55	**0.99**	0.94	0.73	0.55
	FDR:	0.18	0.24	0.38	0.72	0.02	0.05	0.42	0.66	0.01	0.08	0.4	0.66
	TPR:	0.98	0.83	1	0.63	0.94	0.83	0.96	0.99	0.94	0.85	0.97	0.97
20%	MCC:	**0.8**	0.52	0.74	0.28	**0.88**	0.75	0.64	0.52	**0.92**	0.78	0.66	0.5
	FDR:	0.24	0.51	0.42	0.81	0.15	0.29	0.53	0.69	0.08	0.26	0.51	0.71

case:		D				E			
contamination		SS_W	SS	RLARS	MCP	SS_W	SS	RLARS	MCP
	TPR:	0.98	0.94	1	0.99	0.99	0.88	1	0.8
5%	MCC:	**0.98**	0.89	0.73	0.54	**0.87**	0.67	0.77	0.42
	FDR:	0.01	0.13	0.43	0.67	0.16	0.36	0.37	0.72
	TPR:	0.95	0.84	0.98	0.89	0.98	0.81	0.99	0.68
10%	MCC:	**0.95**	0.71	0.69	0.45	**0.8**	0.53	0.73	0.31
	FDR:	0.05	0.36	0.48	0.73	0.25	0.53	0.43	0.79
	TPR:	0.89	0.77	0.92	0.7	0.96	0.74	0.98	0.49
20%	MCC:	**0.85**	0.52	0.61	0.3	0.62	0.28	**0.68**	0.17
	FDR:	0.18	0.58	0.55	0.81	0.47	0.78	0.5	0.87

Table 3.4: Comparing the performance of Spike-and-Slab Winsorized mean, Spike-and-Slab, RLARS, and minimax concave penalty (MCP) under different contamination scenarios (5%, 10%, and 20%) simulating outliers. Each scenario is simulated 100 times, with the results averaged.

In Table 3.4, we use five different methods for data contamination to simulate outliers and compare the performance of the improved method $\mathbf{SS_W}$ with Spike-and-Slab, RLARS, and minimax concave penalty (MCP, Zhang (2010)). MCP is a non-robust method, and therefore its results highlight the difficulties of variable selection in these five contamination scenarios with varied degrees of severity. The results of $\mathbf{SS_W}$ show that it combines the advantages of Spike-and-Slab and RLARS. In circumstances with low contamination levels, Spike-and-Slab performs well, but $\mathbf{SS_W}$ performs even better. As the contamination levels rise, $\mathbf{SS_W}$ not only outperforms Spike-and-Slab, but also has a lower performance deterioration than the original approaches. In heavy contamination conditions, $\mathbf{SS_W}$ and RLARS achieve similar TPR. The FDR for $\mathbf{SS_W}$ is consistently lower than that of RLARS,

doi:10.6342/NTU202402347

sometimes significantly so. As a result, in most cases, the MCC of **SS$_W$** is also higher than that of RLARS.

Comparison for Robustness Against High Leverage Points

We previously compared cases containing outliers. Now we will simulate scenarios with high leverage points. The parameter settings remain as before. We consider different distribution scenarios to contaminate the data and simulate high leverage points:

- I. $X_{ij} \sim (1-\alpha)\mathcal{N}(0,1) + \alpha\mathcal{N}(50,1)$, high leverage contamination;

- II. $\epsilon \sim (1-\alpha)\mathcal{N}(0,1) + \alpha t(2)$ and $X_{ij} \sim \mathcal{N}(5,1)$, heavy-tailed t-distribution and high leverage contamination;

where α is the proportion of contaminated data.

In Table 3.5, we simulated high leverage points (case I) and the simultaneous presence of heavy tails and high leverage points (case II). It is obvious that RLARS performs poorly in case I, whereas Spike-and-Slab works extremely well. **SS$_W$** also performs excellently, unaffected by RLARS. In case II, RLARS maintains its robust characteristics; Spike-and-Slab performs well; however, **SS$_W$** performs even better.

doi:10.6342/NTU202402347

case:		**I**				**II**			
contamination		**SS_W**	**SS**	**RLARS**	**MCP**	**SS_W**	**SS**	**RLARS**	**MCP**
	TPR:	1	1	0.02	1	1	1	1	1
5%	MCC:	**1**	**1**	0.02	**1**	**1**	0.94	0.77	0.64
	FDR:	0	0	0.24	0	0	0.1	0.37	0.56
	TPR:	1	1	0.02	1	1	1	1	0.99
10%	MCC:	**1**	**1**	0.05	**1**	**1**	0.94	0.77	0.61
	FDR:	0	0	0.3	0	0	0.11	0.38	0.59
	TPR:	1	1	0.14	1	1	1	1	1
20%	MCC:	**1**	**1**	0.19	**1**	**1**	0.95	0.75	0.6
	FDR:	0	0	0.5	0	0	0.1	0.41	0.61

Table 3.5: Comparing the performance of Spike-and-Slab Winsorized mean, Spike-and-Slab, RLARS, and minimax concave penalty (MCP) under different contamination scenarios (5%, 10%, and 20%) simulating high leverage. Each scenario is simulated 100 times, with the results averaged.

doi:10.6342/NTU202402347

Chapter 4 Real Data Applications

4.1 Gene Expression - Obese Diabetic Mice

We analyze the performance of $\mathbf{SS_W}$ on real data. The first dataset we used was from Lan et al. (2006), and it consisted of liver samples from 60 obese diabetic mice ($n = 60$), 29 male and 31 female. This dataset contains 22,575 gene expression values ($p = 22,575$), with stearoyl-CoA desaturase 1 (SCD1), glycerol-3-phosphate acyltransferase (GPAT) and phosphoenolpyruvate carboxykinase (PEPCK) as the response variables.

First, we use the robust correlation method mentioned in Section 2.4, employing Winsorized data points, to select the top 200 genes most correlated with SCD1 from the 22,575 gene expression values. Then we utilize the MCP approach to count the number and magnitude of non-zero coefficients and $\hat{\sigma} = 0.16$. We generate non-zero β_{true} values { -3, -2.5, -2, -1.5, -1, 1, 1.5, 2, 2.5, 3 }, a total of 10 non-zero β_{true} values, while the remaining 190 β_{true} values are set to 0. We set $\sigma = 0.16$ and simulate linear regression as follows:

$$Y = X\beta_{\text{true}} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.1}$$

where $X_{60\times200}$ represents the matrix of gene expression values for 60 samples and 200 selected genes. From the results in Table 4.1, we can observe that RLARS has slightly

doi:10.6342/NTU202402347

Method	$\mathbf{SS_W}$	SS	RLARS
TPR	0.57	0.59	0.64
MCC	**0.62**	0.47	0.44
FDR	0.27	0.56	0.63

Table 4.1: (Obese Diabetic Mice) We repeated the simulation 100 times. In each iteration, 10 out of 200 variables were randomly selected to be non-zero, while the remaining variables were set to zero. We compared the performance of the Spike-and-Slab Winsorized mean, Spike-and-Slab, and RLARS.

higher TPR and FDR compared to SS. Additionally, $\mathbf{SS_W}$ shows a comparable TPR to the other two methods, but with the lowest FDR and the best MCC. This indicates that $\mathbf{SS_W}$ performs the best among the three methods.

Finally, we used the responses SCD1, GPAT, and PEPCK with the robust correlation filtered $X_{60\times200}$, selecting a total of 9 variables for SCD1, and 2 variables each for GPAT and PEPCK. For the selected indices, please refer to Table 4.2. Based on Table 4.2, the indices selected by the Horseshoe and S5 methods, as derived from Ray and Bhattacharya (2018), indicate that both methods perform similarly well. $\mathbf{SS_W}$ selected more variables for SCD1 compared to the other two methods, but it included some variables that were also chosen by the other methods. However, for GPAT and PEPCK, $\mathbf{SS_W}$ and the other two methods selected a similar number of variables, with no overlap. Overall, $\mathbf{SS_W}$ demonstrates potential in gene selection.

Response	$\mathbf{SS_W}$	Horseshoe	S5
SCD1	**10310**, **4729**, 17961, 10769, 18945, 570, 21634, **6002**, 8341	**6002**, **10310**	4664, **4729**
GPAT	6896, 20017	10854	17498, 18639
PEPCK	6987, 2150	7640, 18623	7640, 18558

Table 4.2: The numbers in the table represent indices, with bold numbers indicating indices that are also selected by SS_W. The indices selected by the Horseshoe and S5 methods are derived from Ray and Bhattacharya (2018).

doi:10.6342/NTU202402347

4.2 Gene Expression - NCI60

We continue to explore the performance of $\mathbf{SS_W}$ on real data. For the second dataset, we used NCI-60 cancer cell data (Reinhold et al., 2012). This collection contains 60 human cancer cell lines representing nine different types of cancers, including breast cancer, renal cancer, prostate cancer, and others. However, one observation is excluded because two samples were eventually recognized as being from the same patient and combined into a single cell line, resulting in 59 observations ($n = 59$). The dataset comprises 22,283 gene expression values ($p = 22,283$). We employed the 92nd protein as the response variable, following the process outlined by Alfons (2021).

Similar to Section 4.1, we used the robust correlation method mentioned in Section 2.4 with winsorized data points to select the top 200 gene expression values out of 22,283 that had the highest correlation with the response variable. We then used MCP to observe the non-zero coefficients and their quantities, and $\hat{\sigma} = 0.78$. We generated the non-zero β_{true} values as {-1, -0.875, -0.75, -0.625, -0.5, 0.5, 0.625, 0.75, 0.875, 1}, resulting in 10 non-zero β_{true} values, with the remaining 190 β_{true} values set to zero. We set $\sigma = 0.78$ and simulate linear regression as follows:

$$Y = X\beta_{\text{true}} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.2}$$

where $X_{59\times200}$ represents the matrix of gene expression values for 59 samples and 200 selected genes.

From the results in Table 4.3, it is clear that RLARS generally performs worse than SS. As mentioned earlier, the $\mathbf{SS_W}$ method uses RLARS to improve SS. The results show

doi:10.6342/NTU202402347

Method	$\mathbf{SS_W}$	SS	RLARS
TPR	0.61	0.81	0.46
MCC	0.62	**0.64**	0.33
FDR	0.32	0.45	0.68

Table 4.3: (NCI60) We repeated the simulation 100 times. In each iteration, 10 out of 200 variables were randomly selected to be non-zero, while the remaining variables were set to zero. We compared the performance of the Spike-and-Slab Winsorized mean, Spike-and-Slab, and RLARS.

that $\mathbf{SS_W}$ has slightly lower TPR and FDR compared to SS, but the MCC for both $\mathbf{SS_W}$ and SS is quite similar. Therefore, even when RLARS performs significantly worse than SS, it does not make $\mathbf{SS_W}$ perform much worse than SS.

Final, we used the response 92nd protein and the robust correlation filtered $X_{59\times100}$, reducing the variables to 100 as done in Alfons (2021), and selecting a total of 15 variables. For the selected indices, please refer to Table 4.4. Based on Table 4.4, the indices selected by the sparseLTS methods are derived from Alfons (2021). The $\mathbf{SS_W}$ technique selected 15 variables, 4 of which overlapped with those selected by the sparseLTS approach. This shows that the $\mathbf{SS_W}$ approach matches the sparseLTS method for some essential variables. The total number of variables selected by the two methods is comparable. Each approach chose distinct variables, showing disparities in variable selection.

Response	$\mathbf{SS_W}$	sparseLTS
92nd protein	**8502**, 1124, **134**, **1106**, **20125**, 2503, 18057, 10193, 8266, 19073, 12980, 8706, 9269, 2030, 2074	**8502**, 21786, **134**, 4454, **1106**, **20125**, 8510, 14785, 17400, 8460, 8120, 18447, 15622, 7696, 5550, 16784, 13547

Table 4.4: The numbers in the table represent indices, with bold numbers indicating indices that are also selected by $\mathbf{SS_W}$. The indices selected by the sparseLTS methods are derived from Alfons (2021).

doi:10.6342/NTU202402347

Chapter 5 Conclusion

We started by comparing the performance of Spike-and-Slab and Horseshoe priors in variable selection in high-dimensional scenarios. There were no significant differences between Spike-and-Slab and Horseshoe at various signal-to-noise ratios (SNR) and sparsity (SP) levels. However, in the presence of data contamination, especially with outliers, Spike-and-Slab outperformed Horseshoe and showed relatively superior robustness. Additionally, Spike-and-Slab has a computing speed that is 5-10 times faster than Horseshoe. Based on these three reasons, we chose Spike-and-Slab and improved its robustness. By adjusting its posterior samples with RLARS, we proposed the Spike-and-Slab Winsorized means ($\mathbf{SS_W}$) method, applying different levels of winsorization to different variables. Simulation results show that $\mathbf{SS_W}$ outperforms both Spike-and-Slab and RLARS in various scenarios, sometimes significantly so. Since $\mathbf{SS_W}$ uses RLARS to adjust Spike-and-Slab, even when RLARS performs worse than Spike-and-Slab, the performance of $\mathbf{SS_W}$ remains comparable to Spike-and-Slab and does not deteriorate significantly. Therefore, $\mathbf{SS_W}$ is a useful and practical method of variable selection in high-dimensional situations.

doi:10.6342/NTU202402347

doi:10.6342/NTU202402347

References

Alfons, A. (2021). robusthd: An r package for robust regression with high-dimensional data. Journal of Open Source Software, 6(67):3786.

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics, pages 226–248.

Bai, R., Ročková, V., and George, E. I. (2021). Spike-and-slab meets lasso: A review of the spike-and-slab lasso. Handbook of Bayesian Variable Selection, pages 81–108.

Bellman, R. E. (1957). Dynamic Programming. Princeton University Press.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. Biometrika, 103(4):985–991.

Biswas, N., Mackey, L., and Meng, X.-L. (2022). Scalable spike-and-slab. In International Conference on Machine Learning, pages 2021–2040. PMLR.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika, 97(2):465–480.

doi:10.6342/NTU202402347

Chen, Y., Du, P., and Wang, Y. (2014). Variable selection in linear models. Wiley Interdisciplinary Reviews: Computational Statistics, 6(1):1–9.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. The Annals of Statistics, 32(2):407–499.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. Annals of Statistics, 36(6):2605–2637.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. National Science Review, 1(2):293–314.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In Sanz-Sole, M., Soria, J., Varona, J. L., and Verdera, J., editors, Proceedings of the International Congress of Mathematicians, pages 595–622.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society Series B: Statistical Methodology, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20(1):101.

Filzmoser, P. and Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. Wiley Interdisciplinary Reviews: Computational Statistics, 13(4):e1524.

doi:10.6342/NTU202402347

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. Statistica Sinica, pages 339–373.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). Model selection in Bayesian neural networks via Horseshoe priors. Journal of Machine Learning Research, 20(182):1–46.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. Journal of the American Statistical Association, 110(509):435–448.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430):773–795.

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. Journal of the American Statistical Association, 102(480):1289–1299.

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. Chemometrics and Intelligent Laboratory Systems, 172:211–222.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F., et al. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genetics, 2(1):e6.

doi:10.6342/NTU202402347

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of bayesian variable selection methods: what, how and which. Bayesian Analysis, 4:85–118.

Park, T. and Casella, G. (2008). The Bayesian LASSO. Journal of the American Statistical Association, 103(482):681–686.

Piironen, J. and Vehtari, A. (2017). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In Artificial Intelligence and Statistics, pages 905–913. PMLR.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. Journal of the American Statistical Association, 92(437):179–191.

Ray, P. and Bhattacharya, A. (2018). Signal adaptive variable selector for the horseshoe prior. arXiv preprint arXiv:1810.09004.

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J., and Pommier, Y. (2012). Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Research, 72(14):3499–3511.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79(388):871–880.

Serneels, S., Croux, C., Filzmoser, P., and Van Espen, P. J. (2005). Partial robust m-regression. Chemometrics and Intelligent Laboratory Systems, 79(1-2):55–64.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288.

doi:10.6342/NTU202402347

Wilcox, R. (2005). Trimming and winsorization. Encyclopedia of Biostatistics, 8.

Wold, H. (1985). Partial least squares. In Kotz, S. and Johnson, N. L., editors, Encyclopedia of Statistical Sciences, volume 6, pages 581–591. John Wiley, New York.

Yohai, V. J. (1987). High breakdown-point and high-efficiency robust estimates for regression. Annals of Statistics, 15(2):642–656.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38(2):894–942.

Zou, H. (2006). The adaptive LASSO and its oracle properties. Journal of the American Statistical Association, 101(476):1418–1429.

doi:10.6342/NTU202402347