

# Stability Selection via Variable Decorrelation

Mahdi Nouraie<sup>1</sup>, Connor Smith<sup>1</sup>, and Samuel Muller<sup>1,\*1,2</sup>

<sup>1</sup>School of Mathematical and Physical Sciences, Macquarie University

<sup>2</sup>School of Mathematics and Statistics, The University of Sydney

## Abstract

The Lasso is a prominent algorithm for variable selection. However, its instability in the presence of correlated variables in the high-dimensional setting is well-documented. Although previous research has attempted to address this issue by modifying the Lasso loss function, this paper introduces an approach that simplifies the data processed by Lasso. We propose that decorrelating variables before applying the Lasso improves the stability of variable selection regardless of the direction of correlation among predictors. Furthermore, we highlight that the irrepresentable condition, which ensures consistency for the Lasso, is satisfied after variable decorrelation under two assumptions. In addition, by noting that the instability of the Lasso is not limited to high-dimensional settings, we demonstrate the effectiveness of the proposed approach for low-dimensional data. Finally, we present empirical results that indicate the efficacy of the proposed method across different variable selection techniques, highlighting its potential for broader application. The **DVS** R package is developed to facilitate the implementation of the methodology proposed in this paper.

Keywords: Bioinformatics, Consistency, Feature Selection, Gram-Schmidt Process, Lasso, QR Factorisation, Stability, Variable Selection

## 1 Introduction

Linear regression is formally defined as  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\mathbf{Y}$  denotes the  $n$ -vector of the response variable,  $\mathbf{X} \in \text{mat}(n, p)$  denotes the fixed design matrix,  $\boldsymbol{\beta}$  denotes the  $p$ -vector of regression coefficients, and  $\boldsymbol{\epsilon}$  denotes the  $n$ -vector of random errors. The intercept term is omitted here because, without loss of generality, we assume that the response vector is centred, that is,  $\sum_{i=1}^n \mathbf{Y}_i = 0$ . The least absolute shrinkage and selection operator [Lasso; Tibshirani, 1996] is a prominent variable selection method that achieves sparsity by adding an  $\ell_1$ -norm to the least-squares loss function. The regression coefficients of the Lasso are estimated through

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left( \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right), \quad (1)$$

where  $\lambda \in \mathbb{R}^+$  denotes the Lasso regularisation parameter.

---

\*Address for correspondence: samuel.muller@mq.edu.au

Variable selection, following the terminology of Meinshausen and Bühlmann [2010], refers to categorising variables into two distinct groups: the signal group  $S := \{k \neq 0 \mid \beta_k \neq 0\}$  and the noise group  $N := \{k \neq 0 \mid \beta_k = 0\}$ , where  $S \cap N = \emptyset$ . The primary objective of variable selection is to accurately identify the signal group  $S$ . When using the Lasso, the set of non-zero coefficients  $\hat{S}(\lambda) := \{k \neq 0 \mid \hat{\beta}_k(\lambda) \neq 0\}$ , can be derived from solving Equation (1) through convex optimisation.

The stability of a variable selection method is related to its ability to consistently identify the same variables across different training sets sampled from the same underlying distribution [Kalousis et al., 2007]. This property serves as an indicator of the reproducibility and generalisability of the selection results and has received significant attention in recent years (see, e.g., Meinshausen and Bühlmann [2010], Nogueira et al. [2018], and Soloff et al. [2024]). In the high-dimensional setting, where  $p \gg n$ , the presence of correlated variables within  $X$  can directly affect the stability of the variable selection carried out by Lasso [Wang et al., 2018]. To be more specific, when two variables are highly correlated and their effect sizes are unconstrained, they exhibit a nearly identical effect on the response variable [Wang et al., 2018]. In such cases, Lasso tends to select at random only one of these variables [Xu et al., 2012, Wang et al., 2018]. In general, if an irrelevant variable is highly correlated with the relevant variables, Lasso may not be able to distinguish it from the relevant variables with any amount of data and any amount of regularisation [Leng et al., 2006, Zhao and Yu, 2006].

In this paper, we propose to decorrelate predictor variables before performing variable selection by Lasso. The proposed variable decorrelation process transforms the predictors into an orthonormal basis aiming to reduce noise from inter-variable dependencies and improve interpretability. We show that this transformation improves the stability of variable selection for the Lasso. Although the orthonormal design matrix has long been a subject of interest (see, e.g., Tibshirani [1996] and Yuan and Lin [2006]), to the best of our knowledge, this is the first paper to propose addressing the instability of the Lasso through an orthonormalisation process. We further show that the irrepresentable condition [Zhao and Yu, 2006, Meinshausen and Bühlmann, 2006], which guarantees consistent variable selection by the Lasso, holds after variable decorrelation under two assumptions. In addition, we highlight that the instability of the Lasso is not confined to high-dimensional settings, and variable decorrelation can serve as an effective strategy for achieving stable variable selection in low-dimensional data as well.

Several previous studies have attempted to improve the stability of the Lasso by modifying its loss function to better handle correlated variables (see, e.g., Chen et al. [2013], Takada et al. [2018], and Wang et al. [2018]). A complementary approach, stability selection, enhances the stability of variable selection by applying a selection method like the Lasso to multiple random half-samples of the original dataset [Meinshausen and Bühlmann, 2010, Shah and Samworth, 2013]. Stability selection identifies relevant variables by evaluating their selection frequencies across various randomly drawn sub-samples of observations. In this paper, we use the latter approach to identify stable variables following the decorrelation of the variables.

The Gram–Schmidt process [Björck, 1994] is a fundamental technique for generating an orthonormal set of vectors that spans the same subspace as the original set. The Gram–Schmidt process is traditionally used to compute the thin QR factorisation [Golub and Van Loan, 2013], and in recent years, has been applied to variable ranking, screening, and selection problems (see,

e.g., Stoppiglia et al. [2003], Liu et al. [2018], Wang et al. [2020], and Skogholt et al. [2023]). The Gram–Schmidt process is presented in Algorithm 2 in the appendix. The resulting matrix  $Q$  in Algorithm 2 has the same dimensions as the input matrix  $X$ , with its columns mutually orthogonal and normalised to have an  $\ell_2$ -norm of one. In this paper, we show that using  $Q$ , obtained by applying Algorithm 2 to the design matrix  $X$ , as a surrogate for  $X$  in variable selection results in greater stability; however, we first comment on its limitations and then focus on refinements.

Algorithm 2 sequentially orthogonalises columns of  $X$  from index 1 to  $p$ , giving earlier variables greater influence over the resulting basis when correlations are non-zero. This creates a competitive dynamic, where later correlated variables must offer additional information to be represented. Here, we present a toy example to illustrate the index sensitivity of Algorithm 2. The dataset contains  $n = 100$  samples and  $p = 5$  variables, with  $X_2 = 0.8X_1$  introducing correlation. The response  $\mathbf{Y}$  is generated from  $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^\top$  with standard normal noise. We apply the Lasso to the orthonormal matrix  $Q$  in the two cases: once obtained from the original design matrix, and once from the reordered design matrix after swapping  $X_1$  and  $X_2$ . In both cases, only the first coefficient is non-zero, highlighting how the order of the variables affects the selection results. The corresponding implementation code is available on GitHub<sup>1</sup>. Section 4 presents further simulations showing that applying Lasso to  $Q$  often yields poor selection accuracy. To address this, we propose reordering columns before applying the Gram–Schmidt process. Therefore, in this paper, we introduce a method which involves three key steps: (i) reordering the variables, (ii) applying Gram-Schmidt orthogonalisation, and (iii) performing variable selection on  $Q$  using the stability selection with the Lasso.

The rest of the paper is organised as follows. Section 2 presents and justifies our method. Section 3 describes the synthetic and real datasets used in the paper. Section 4 presents the numerical results, and Section 5 discusses related work and situates the proposed approach within its broader context and potential impact. The appendix contains the Gram-Schmidt process algorithm, proofs of the corollaries, and supplementary numerical results.

## 2 Methodology

In this section, we propose a method that reorders the columns of  $X$ , decorrelates the variables by projecting them onto an orthonormal basis, and integrates the resulting orthonormal matrix into the stability selection framework to improve the selection stability of the Lasso in the presence of correlated predictors.

We consider a dataset  $\mathcal{D} := \{(\mathbf{x}_i^\top, y_i)\}_{i=1}^n$  where each element of  $\mathcal{D}$  consists of a univariate response  $y_i \in \mathbb{R}$  and a  $p$ -dimensional fixed variable vector  $\mathbf{x}_i^\top \in \mathbb{R}^p$ . We assume that the columns of the design matrix  $X$  are centred and scaled by their means and standard deviations. Following the basic assumptions of linear regression, it is also assumed that  $E(\mathbf{Y} | X)$  is linear in the coefficients,  $E(\boldsymbol{\epsilon} | X) = 0$ ,  $\text{Var}(\boldsymbol{\epsilon} | X) = \sigma^2 < \infty$ , and error terms are uncorrelated with predictor variables. In addition, we assume that the columns of  $X$  are not linearly dependent on each other and that the components of  $\boldsymbol{\epsilon}$  are independent of each other.

---

<sup>1</sup>[https://github.com/MahdiNouraie/Variable-Decorrelation/blob/main/Code/Toy\\_example.R](https://github.com/MahdiNouraie/Variable-Decorrelation/blob/main/Code/Toy_example.R)

Given that Algorithm 2 is sensitive to the order of column indices, we propose that the columns of the design matrix first be sorted according to their predictive power for the response variable. To achieve this, we advocate for the use of sure screening methods, given their favourable asymptotic properties, as an appropriate approach for ordering the variables prior to the decorrelation and selection process. Specifically, we employ Air-Holp [Joudah et al., 2025], a data-adaptive variant of ridge high-dimensional ordinary least-squares projection [Ridge-HOLP; Wang and Leng, 2015]. Although other screening and ranking methods may be used, comparing the efficacy of different ranking methods is beyond the scope of this paper. The variables should be ordered according to the output of the chosen ranking method, ensuring that the variable with the highest predictive power for the response variable is placed in the first column, etc. In the next step, we decorrelate the variables within the ordered design matrix by projecting them onto a new space of orthonormal vectors by Algorithm 2. Then, we propose using  $Q$  and  $\mathbf{Y}$  in the stability selection using the Lasso to select the relevant variables. The general method is outlined in Algorithm 1.

---

#### Algorithm 1 Stability Selection via Variable Decorrelation

---

- 1: **Input:** Design matrix  $X$ , response vector  $\mathbf{Y}$
  - 2: **Output:** Relevant variables
  - 3: **Step 1:** Sort the columns of  $X$  using Air-HOLP to prioritise variables with higher predictive power for the response variable.
  - 4: **Step 2:** Project the ordered design matrix  $X$  onto a new space  $Q$  with orthonormal vectors using Algorithm 2 to decorrelate the variables.
  - 5: **Step 3:** Integrate the transformed design matrix  $Q$  and the response vector  $\mathbf{Y}$  into the stability selection framework using the Lasso.
- 

It should be noted that by applying the Lasso to the orthonormal matrix  $Q$ , the resulting solution has a closed-form expression given by  $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \lambda)^+$ , where  $\hat{\beta}_j^0$  denotes the full least-squares estimate [Tibshirani, 1996].

## Justifying the Surrogate

### Why columns of $Q$ are good proxies for columns of $X$ ?

As shown in Algorithm 2, the  $j$ th column of  $R$  contains the normalising factor for the corresponding orthogonal vector in  $Q$  on the diagonal, with zeros below it. The entries above the diagonal capture the linear dependencies between  $X_j$  and the previously orthonormalised vectors. Therefore, we conclude that the component of  $X_j$  contributing uniquely to the span of  $X$  is retained in its orthonormalised form as  $Q_j$ . This enables us to use  $Q$  as a suitable surrogate for variable selection compared to  $X$  when the variables are correlated.

### Is the principle of resampling from observations preserved when using $Q$ in stability selection?

Stability selection relies on repeatedly drawing random sub-samples of the observations to enhance the stability of variable selection. An important consideration in the context of stability

selection is whether it is valid to use the orthonormal matrix  $Q$  within the stability selection, given that  $Q$  is computed using all observations from the original design matrix  $X$ ?

For any sub-sample of observations, let  $X_{b\bullet}$  represent the corresponding rows of  $X$ . Since  $X = QR$ , we have  $X_{b\bullet} = Q_{b\bullet}R$ . This relationship implies that sub-sampling the rows of  $Q$  is intrinsically linked to sub-sampling the original observations in  $X$ , followed by a fixed transformation determined by  $R$ . Consequently, the fundamental principle of stability selection over sub-samples of the original data is preserved.

## Regularisation Tuning

In our implementations, we follow the guidelines recently introduced by Nouraie and Muller [2024], who proposed a method for tuning the Lasso regularisation parameter that balances selection stability and prediction accuracy, consistent with the approach proposed by Sun et al. [2013]. Specifically, Nouraie and Muller [2024] recommended recording the selection results for each regularisation value  $\lambda$  in the corresponding binary selection matrix  $M(\lambda) \in \text{mat}(B, p)$ , where  $B$  denotes the number of sub-samples used when applying the Lasso in the stability selection process. They leveraged the stability measure introduced by Nogueira et al. [2018] to assess the stability of  $M(\lambda)$  and determine the optimal regularisation value. The stability measure is defined as

$$\hat{\Phi}(M(\lambda)) := 1 - \frac{\frac{1}{p} \sum_{j=1}^p s_j^2}{\frac{q(\lambda)}{p} \left(1 - \frac{q(\lambda)}{p}\right)}; \quad \lambda \in \Lambda, \quad (2)$$

where  $s_j^2$  denotes the unbiased sample variance of the binary selection statuses of the  $j$ th variable within  $M(\lambda)$ ,  $q(\lambda)$  denotes the average number of variables selected under the regularisation parameter  $\lambda$ , and  $\Lambda$  denotes the set of candidate regularisation values. Nouraie and Muller [2024] proposed  $\lambda_{\text{stable}}$ , defined as the smallest regularisation value for which the corresponding stability exceeds 0.75, and  $\lambda_{\text{stable-1sd}}$ , defined as the smallest regularisation value for which the corresponding stability remains within one standard deviation of the maximum stability achieved.  $\lambda_{\text{stable}}$  is identified as an optimal choice on the Pareto front [Pareto, 1896] of selection stability and prediction accuracy, while  $\lambda_{\text{stable-1sd}}$  is recommended when  $\lambda_{\text{stable}}$  is not attainable.

## Consistent Variable Selection

Zhao and Yu [2006] and Meinshausen and Bühlmann [2006] introduced the irrepresentable condition for the Lasso, which ensures the consistent variable selection. This condition can be mathematically formulated as  $\|X_N^\top X_S (X_S^\top X_S)^{-1}\|_\infty < 1$ , where  $X_N := [X_j \mid j \in N]$  represents the sub-matrix of  $X$  that contains only noise variables, and  $X_S := [X_j \mid j \in S]$  is the sub-matrix that contains only signal variables. Here, we demonstrate that by orthonormalising the design matrix, this condition is satisfied under two assumptions.

**Corollary 1.** *If the columns of the design matrix  $X$  are not linearly dependent on each other, and if the variable ordering method positions all relevant variables in earlier columns relative to the irrelevant variables that are correlated with them, the irrepresentable condition for the*

Lasso is satisfied by employing the matrix  $Q$ , which represents the orthonormal matrix obtained through applying Algorithm 2 to the ordered  $X$ , as the new design matrix.

The proof of Corollary 1 is shown in the appendix. The first assumption in Corollary 1 ensures that division by zero is avoided during the normalisation step in Algorithm 2. The second assumption pertains to the effectiveness of the ranking method employed. For Ridge-HOLP, it is established in Wang and Leng [2015] that the method is screening consistent—that is, if a model of the same size as the true model is retained, it asymptotically coincides with the true model. However, the requirement in Corollary 1 is weaker. We do not require that all relevant variables precede all irrelevant ones in the ranking; rather, it suffices that each relevant variable precedes those irrelevant variables that are correlated with it. According to Wang and Leng [2015], a critical condition for the screening consistency of Ridge-HOLP is that the condition number of the covariance matrix of the design matrix satisfies  $\kappa(\Sigma) \leq cn^\tau$ , where  $c > 0$  and  $0 \leq \tau < 1/7.5$ . Here,  $\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$  denotes the condition number of  $\Sigma$ , and  $\lambda_{\max}(\Sigma)$  and  $\lambda_{\min}(\Sigma)$  denote the largest and smallest eigenvalues of  $\Sigma$  respectively. One instance in which this condition is met is when the groups of correlated variables are mutually independent, each group contains finitely many variables, and the variables within each group exhibit a compound symmetric structure with a correlation value less than one. The assumptions required for the screening consistency of Ridge-HOLP are stated in Theorem 3 of Wang and Leng [2015].

## Stable Variable Selection

Although the instability of Lasso in the presence of correlated variables has been recognised for years, Faletto and Bien [2022] formally demonstrated that in the context of a simple linear model with two highly correlated, equally relevant variables and one less important variable, as  $n$  approaches infinity, the selection frequencies of the two more relevant variables tend to 0.5, while the selection frequency of the less relevant variable tends to 1 when stability selection is applied using Lasso. Building on this, we now demonstrate how our method improves selection stability.

**Corollary 2.** *Assume that under the conditions specified in Theorem 1 and Corollary 2 of Faletto and Bien [2022], within the design matrix  $X \in \text{mat}(n, 3)$ , there exist two highly correlated and equally relevant variables, denoted  $X_1$  and  $X_2$ , which are uncorrelated with  $X_3$ . Let the corresponding regression coefficients be  $\beta_1 = \beta_2 > \beta_3 = 1$ . When stability selection with Lasso is applied to the orthonormal matrix  $Q$ , obtained by applying Algorithm 2 to  $X$ , it results in higher selection stability compared to applying stability selection with Lasso directly to  $X$ . This improvement in stability is evident when the stability measure defined in Equation (2) is used to evaluate stability.*

The proof of Corollary 2 is shown in the appendix. As is evident, the ordering step is not involved in Corollary 2. As shown in Corollary 1, this step affects selection consistency rather than stability. In addition, since Corollary 2 is established in a low-dimensional setting, we conclude that instability, although often discussed in high-dimensional contexts, is also a concern in low-dimensional settings due to the role of correlation. The numerical results

supporting this conclusion are provided in Section 4. It should also be noted that, as argued by Nogueira et al. [2018], any proper stability measure should be a strictly decreasing function of the variance of the selection status of the variables. Therefore, Corollary 2 holds for any proper stability measure, not only for the specific measure defined in Equation (2).

### 3 Datasets

We evaluated our method through experiments on both synthetic and real bioinformatics datasets, as presented below.

#### Synthetic Data

As synthetic data, we consider a dataset with a sample size of  $n = 50$  that includes  $p = 500$  predictor variables. For each dataset, the predictor variables  $\mathbf{x}_i^\top$  are independently drawn from the Normal distribution  $\mathcal{N}(\mathbf{0}, \Sigma)$ , where  $\Sigma \in \text{mat}(p, p)$  has a diagonal of ones, and the off-diagonal elements are defined as follows

$$\sigma_{jk} = \begin{cases} \rho_g, & \text{if } j, k \text{ belong to the same group } G_g, \text{ where } g = 1, \dots, 5, \\ 0, & \text{otherwise;} \end{cases}$$

where each group consists of consecutive indices

$$G_g = \left\{ \frac{(g-1)p}{5} + 1, \dots, \frac{gp}{5} \right\}, \quad g = 1, \dots, 5.$$

Each group  $G_g$  is associated with a distinct covariance parameter  $\rho_g$ . Specifically, the covariance values are set as  $\rho_1 = 0.2$ ,  $\rho_2 = 0.4$ ,  $\rho_3 = 0.6$ ,  $\rho_4 = 0.8$ , and  $\rho_5 = 0.9$ . Within each group, a single variable is designated as active, and we consider two extreme cases: (i) the lowest-indexed variable in each group is active, and (ii) the highest-indexed variable in each group is active.

The true regression coefficients of active variables are defined as  $\boldsymbol{\beta} = (4, 3.5, 3, 2.5, 2)^\top$  following the order of the groups and there are 99 inactive variables in each group. The error term  $\boldsymbol{\epsilon}$  is an i.i.d. sample from the standard Normal distribution  $\mathcal{N}(0, 1)$ .

In Section 4, we extend the initial data generation setting to explore two additional scenarios aimed at deepening our understanding of the behaviour of the selection method. First, we investigate the impact of altering the sign of the covariance values to assess how the procedure responds to changes in the direction of the variable associations. Second, we examine a reduced-dimensional setting in which the number of variables is decreased from 500 to 20, while maintaining the original group structure, to evaluate the performance of our method in a low-dimensional context.

#### Riboflavin Data

For the first real example, we use the well-established ‘Riboflavin’ dataset, which focusses on the production of riboflavin (vitamin B2) from various *Bacillus Subtilis*. This dataset, provided

by the Dutch State Mines Nutritional Products, is accessible via `hdi` R package [Dezeure et al., 2015]. It comprises a single continuous response variable that represents the logarithm of the riboflavin production rate, alongside  $p = 4,088$  variables, which correspond to the logarithm of expression levels for 4,088 bacterial genes. The primary objective of analysing this dataset is to identify genes that are associated with riboflavin production, with the ultimate goal of genetically engineering bacteria to enhance riboflavin yield. Data were collected from  $n = 71$  relatively homogeneous samples, which were repeatedly hybridised during a feed-batch fermentation process involving different engineered strains and varying fermentation conditions. Bühlmann et al. [2014] employed stability selection with Lasso and identified three genes—`LYSC_at`, `YOAB_at`, and `YXLD_at`—as relevant genes. Later Nouraie and Muller [2024] examined this dataset using stability selection and Lasso and showed that the stability of results is very low since  $\lambda_{\text{stable}}$  does not exist and  $\hat{\Phi}(M(\lambda_{\text{stable}-1\text{sd}}))$  converges to around 0.2.

## Affymetrix Rat Genome 230 2.0 Array

As an additional real-world example, we investigate ‘Affymetrix Rat Genome 230 2.0 Array’ microarray data introduced by Scheetz et al. [2006]. This dataset comprises  $n = 120$  twelve-week-old male rats, with expression levels recorded for nearly 32,000 gene probes for each rat. The primary objective of this analysis is to identify the probes most strongly associated with the expression level of the TRIM32 probe (`1389163_at`), which has been linked to the development of Bardet-Biedl syndrome [Chiang et al., 2006]. This genetically heterogeneous disorder affects multiple organ systems, including the retina. In accordance with the pre-processing steps outlined by Huang et al. [2008], we excluded gene probes with a maximum expression level below the 25th percentile and those exhibiting an expression range smaller than 2. This filtering process yielded a refined set of  $p = 3,083$  gene probes that demonstrated sufficient expression and variability for further analysis. Nouraie and Muller [2024] analysed this dataset using stability selection and Lasso and showed that the stability of the results is very low since  $\lambda_{\text{stable}}$  does not exist and  $\hat{\Phi}(M(\lambda_{\text{stable}-1\text{sd}}))$  converges to around 0.15.

## 4 Numerical Results

In this section, we illustrate the efficacy of our method through its application to synthetic and real data introduced in Section 3. We used the corresponding GitHub repositories of Joudah et al. [2025] and Nogueira et al. [2018] to execute Air-HOLP and apply the stability measure of Equation (2) respectively. For regularisation tuning, we used the GitHub repository associated with Nouraie and Muller [2024]. To create a set of candidate regularisation values  $\Lambda$ , we use a 10-fold cross-validation using the `cv.glmnet` function from the `glmnet` R package [Friedman et al., 2010] on the entire dataset  $\mathcal{D}$ . The Air-HOLP method takes an input parameter, `Threshold`, which specifies the number of variables to retain during the screening process. In all experiments, this parameter was fixed at 10 to eliminate its influence on the analysis. In addition, the maximum number of iterations for Air-HOLP was set to 10 across all experiments. To implement the Gram-Schmidt process, we used the `grahm_schimdtR` function, which was

sourced from a solution shared on the Stack Overflow platform<sup>2</sup>.

## Synthetic Data

We generate 100 datasets with varying random seeds, following the setting described in Section 3. For each dataset we set the number of sub-samples  $B = 100$ . We keep this setting of number of datasets and number of sub-samples over all synthetic experiments. In our implementation, we use  $\lambda_{\text{stable-lsd}}$  for each dataset, as  $\lambda_{\text{stable}}$  may not exist for some datasets. This approach ensures consistent regularisation tuning across all datasets.

Figure 1 presents a comparative analysis of stability selection using the Lasso, evaluating both selection stability and F1-score, which is the harmonic mean of precision and recall, with and without the implementation of our method. Here, the F1-score is computed to assess selection accuracy, treating the selection task as a binary classification problem. Figures 1a and 1b correspond to the first simulation scenario, whereas Figures 1c and 1d pertain to the second simulation scenario. As illustrated in Figure 1a, the application of the pre-processing steps significantly enhances stability. Figure 1b depicts the average F1-score of variable selection across different values of the stability selection decision making threshold  $\pi_{\text{thr}}$ . Figure 1b reveals that, although the primary focus of this paper is on improving stability, our method also yield a notable enhancement in the F1-score. This observation aligns with expectations in light of the asymptotic consistency established in Corollary 1.

Figure 1c demonstrates a significant increase in stability, while Figure 1d illustrates a higher average F1-score when employing our method across most of the decision threshold values. This outcome is consistent with the results observed in Figure 1a and 1b. Figures 1a and 1c illustrate that a significant enhancement in stability is observed when correlation is removed from the data. In fact, it is difficult to avoid the conclusion that correlation among predictor variables constitutes a primary source of instability in variable selection.

## Why the Ordering Step is Important?

Based on Corollary 1, the ordering step is essential for accurate (consistent) selection. To highlight its importance, we replicate the process for the second simulation scenario, without ordering the variables first. Therefore, we only apply Gram-Schmidt orthonormalisation to the design matrix  $X$ . Figure 2a shows that this improves selection stability, in line with Corollary 2; but Figure 2b reveals that the F1-score remains zero at all thresholds. This demonstrates that without ordering, the results may be stable but inaccurate. Therefore, it can be concluded that the ordering step enhances selection accuracy, while the orthonormalisation step contributes to selection stability.

---

<sup>2</sup><https://stackoverflow.com/questions/15584221/gram-schmidt-with-r>

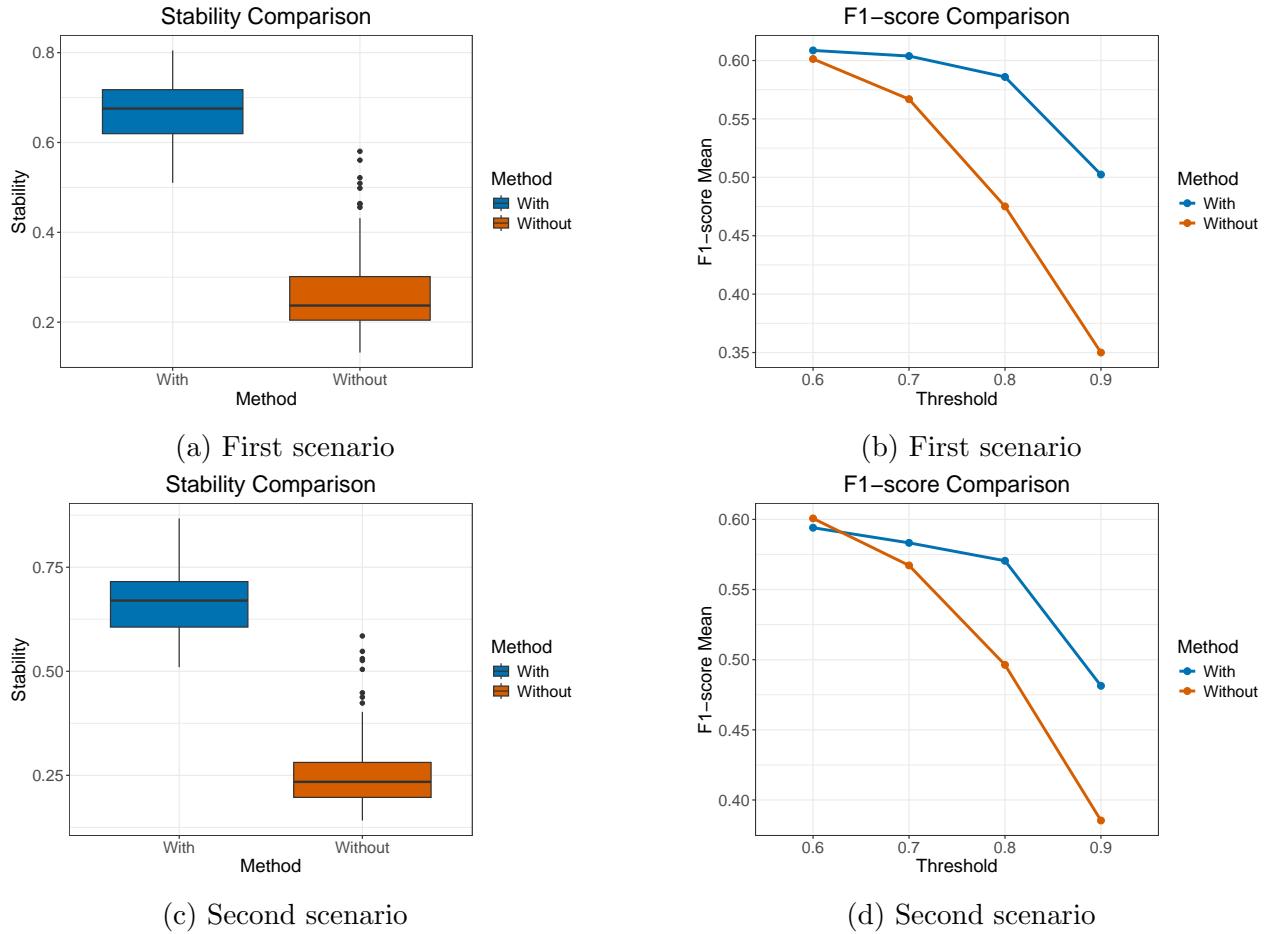


Figure 1: Comparing Stability and F1-score between stability selection using the Lasso with and without the variable decorrelation for simulation scenarios.

## Varying Covariance Signs

To evaluate the effectiveness of our method in relation to the sign of correlations among variables, we consider the second simulation scenario. However, we alter the sign of the covariance values within correlated groups, specifically setting  $\rho_1 = -0.2$ ,  $\rho_2 = -0.4$ ,  $\rho_3 = -0.6$ ,  $\rho_4 = -0.8$ , and  $\rho_5 = -0.9$ . We use the `make.positive.definite` function from the `corpcor` R package [Schafer et al., 2021] to obtain the nearest positive definite matrix, which is then used as the covariance matrix. Figure 3a and 3b demonstrate that, although the signs of the covariances are reversed, the improvements in both stability and accuracy are maintained.

Figure 3c and 3d show the results regarding a more realistic setting where there is a mix of positive and negative covariance values; specifically,  $\rho_1 = 0.2$ ,  $\rho_2 = -0.4$ ,  $\rho_3 = 0.6$ ,  $\rho_4 = -0.8$ ,  $\rho_5 = 0.9$ . Once again, the improvements are retained. In summary, the results from the simulation scenarios demonstrate that our method enhances both selection stability and accuracy when applying the Lasso to correlated high-dimensional data, regardless of the sign of covariance values. This result is consistent with expectations, as no constraints were imposed on the signs of the covariance values in Section 2.

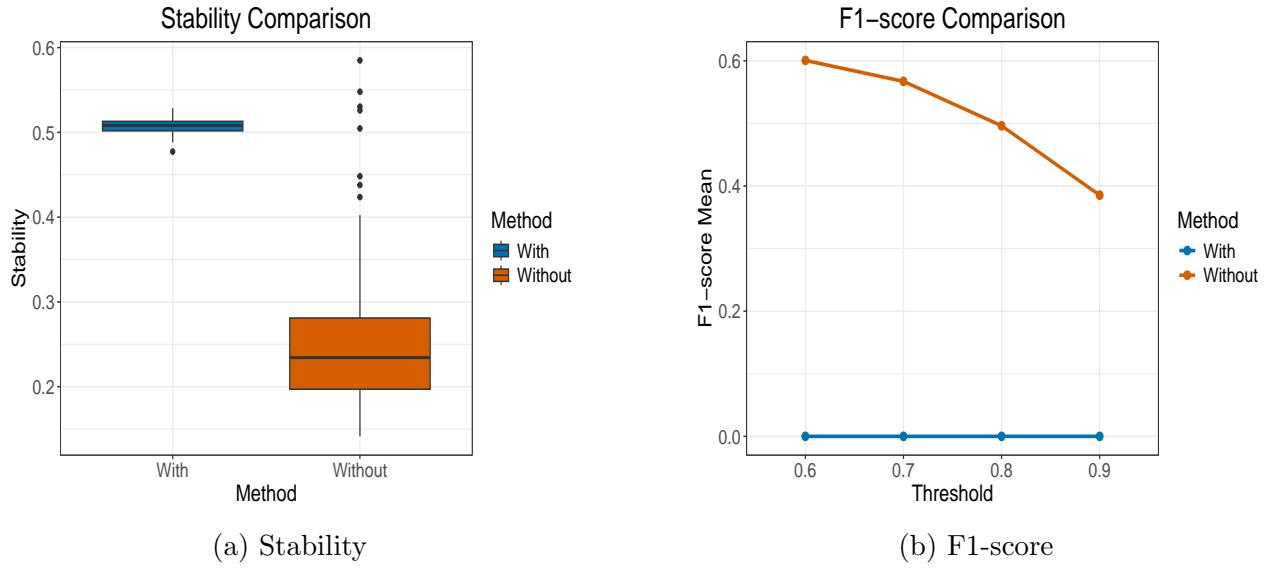


Figure 2: Comparing Stability and F1-score between stability selection using the Lasso with and without the Gram-Schmidt orthonormalisation for the second simulation scenario

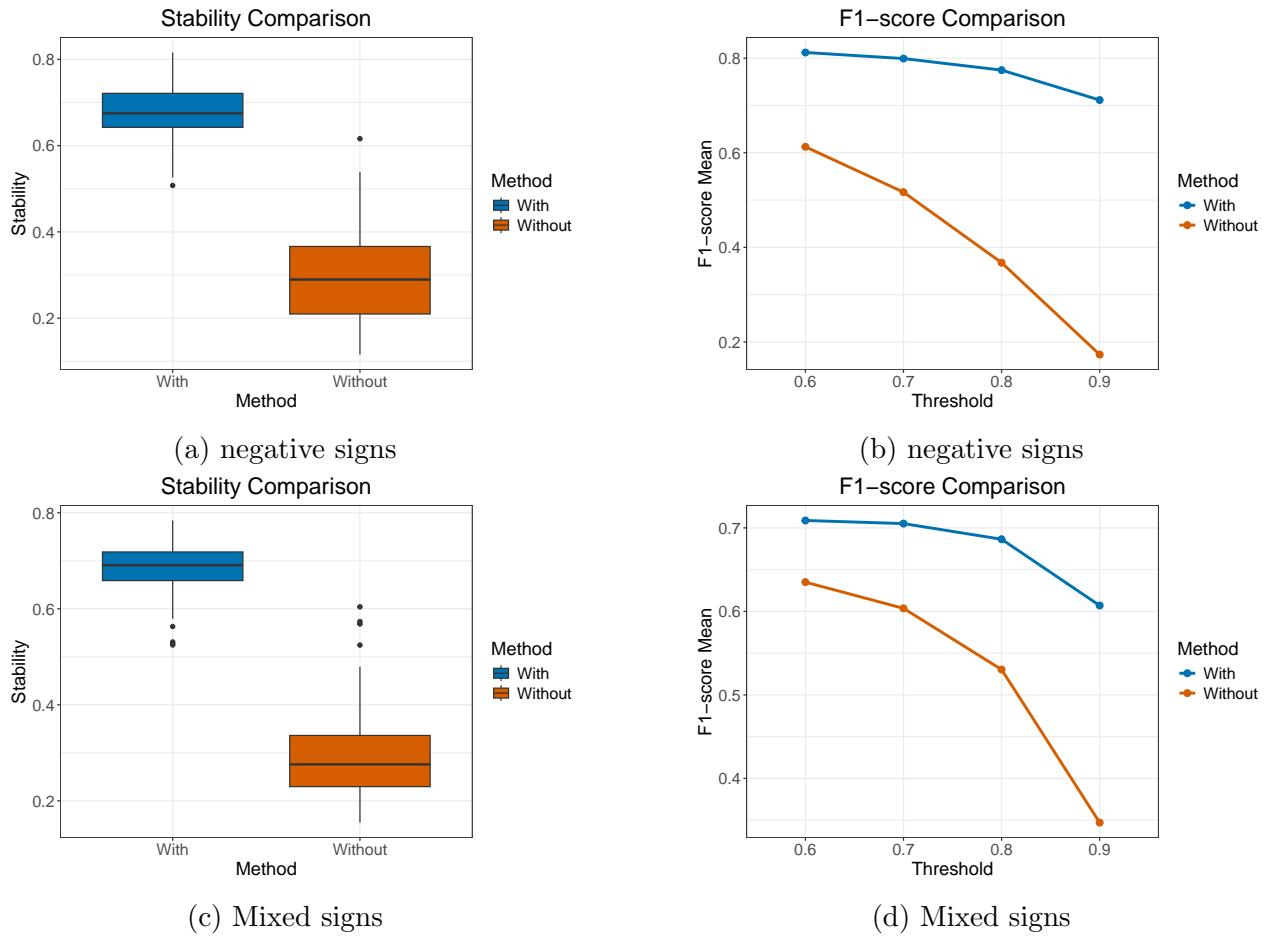


Figure 3: Comparing Stability and F1-score between stability selection using the Lasso with and without the variable decorrelation for the second simulation scenario with varying covariance signs

## Low-Dimensional Data

As discussed in Section 2, Corollary 2 demonstrates that our method improves selection stability even in low-dimensional settings. To further assess this claim, we modify the data generation process in the second simulation scenario by reducing the number of variables from 500 to 20, while keeping all other conditions unchanged. This adjustment places us in a low-dimensional context comprising five groups of variables, each consisting of four variables. Within each group, the variable with the highest index is designated as the relevant one.

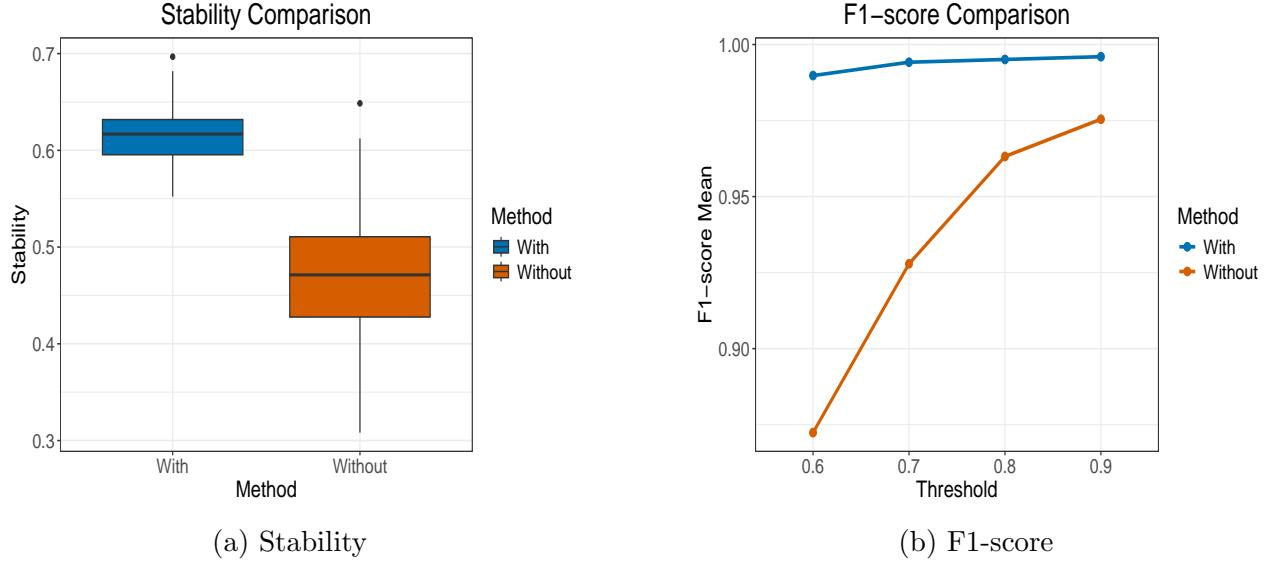


Figure 4: Comparing Stability and F1-score between stability selection using the Lasso with and without the variable decorrelation for the low-dimensional data

Figure 4a illustrates that, as expected, our method leads to an improvement in the selection stability. In addition, Figure 4b shows that the variable decorrelation also leads to an improvement in the selection accuracy. The increasing trend in selection accuracy with respect to  $\pi_{\text{thr}}$  is in contrast to the patterns observed in the previous figures, highlighting a distinctive behaviour in the low-dimensional setting. In addition, across all figures presented so far on the selection accuracy, it is evident that the selection accuracy obtained using variable decorrelation exhibits greater robustness to changes in  $\pi_{\text{thr}}$ .

## Comparison with Alternative Techniques

In this section, we compare our method with several regularisation-based variable selection techniques, including Elastic Net [ENet; Zou and Hastie, 2005], Smoothly Clipped Absolute Deviation [SCAD; Fan and Li, 2001], Minimax Concave Penalty [MCP; Zhang, 2010], and Sparse Group Lasso [SGL; Simon et al., 2013]. The comparison is conducted under the second simulation scenario. The regularisation parameters were selected using the  $\lambda_{\text{stable-1sd}}$  strategy. For the ENet, the mixing parameter `alpha` was set to 0.2. As the SGL requires pre-defined group indices for the variables, the grouping information was provided accordingly.

Figure 5 illustrates that Lasso, when combined with our method, outperforms all other candidates in terms of selection stability. Similarly, Figure 6 demonstrates that our method achieves superior performance across all decision thresholds with respect to the F1-score. These

findings suggest the potential of our method to enhance selection performance not only for the Lasso but possibly for other variable selection methods as well. Although a full exploration of this possibility lies beyond the scope of this paper, supporting numerical results are provided in the appendix.

## Riboflavin Data

As described in Section 3, the Riboflavin dataset consists of  $p = 4,088$  genes, with the primary objective being to identify the key genes that are associated with the riboflavin production. We set the number of sub-samples to  $B = 200$ . In contrast to Nouraie and Muller [2024], the application of our method resulted in the existence of  $\lambda_{\text{stable}}$ . Figure 7a, introduced in Nouraie and Muller [2024], indicates that  $\lambda_{\text{stable}}$  exists and its corresponding stability values have converged to 0.75 after almost 50 sub-samples. The blue shade around the stability line presents the 95% confidence interval for  $\hat{\Phi}(\mathbf{M}(\lambda_{\text{stable}}))$  values. Under the regularisation value  $\lambda_{\text{stable}} = 0.351$  rounded to three decimal places, the genes YXLD\_at and LYSC\_at, with selection frequencies of 0.945 and 0.805 respectively, were identified as relevant genes.

## Affymetrix Rat Genome 230 2.0 Array

As mentioned in Section 3, the rat microarray data consists of  $p = 3,083$  gene probes. The main aim for this data is to identify probes that are associated with the TRIM32 probe. Again, we choose the number of sub-samples  $B = 200$ . Due to the instability of the outcomes,  $\lambda_{\text{stable}}$  does not exist in this problem but, the stability level is much more than what was reported in Nouraie and Muller [2024]. Figure 7b shows that the stability values for  $\lambda_{\text{stable-1sd}}$  are converged to around 0.63. Under the regularisation value  $\lambda_{\text{stable-1sd}} = 0.037$  rounded to three decimal places, the probes X1389457\_at and X1376747\_at with selection frequencies of 0.945 and 0.875 respectively, were identified as relevant probes.

Although Nouraie and Muller [2024] recommended  $B^* = 200$  as a rule of thumb for the stability convergence when applying stability selection with the Lasso, Figure 7a and 7b demonstrate that, with our method, the stability values may converge more rapidly.

## 5 Discussion and Conclusion

Although the classical Gram-Schmidt algorithm is used in this paper, the modified version [Golub and Van Loan, 2013] has been developed to provide better numerical stability. A detailed comparison of these two variants is beyond the scope of this paper.

Meinshausen and Bühlmann [2010] introduced stability selection not only for variable selection, but also for structure estimation tasks such as graphical modelling. In this paper, we demonstrated that decorrelating variables prior to selection enhances stability; however, when the goal is to estimate dependencies among variables—as in graphical models—such decorrelation is not recommended.

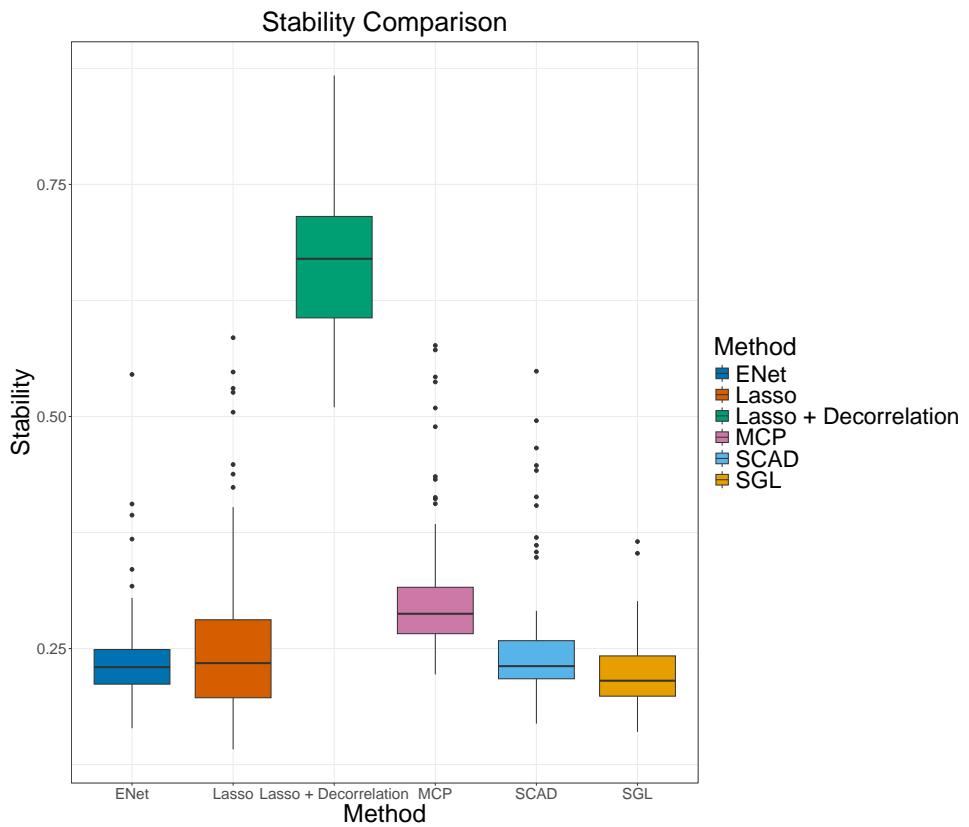


Figure 5: Comparison of selection stability across stability selection with different selection algorithms

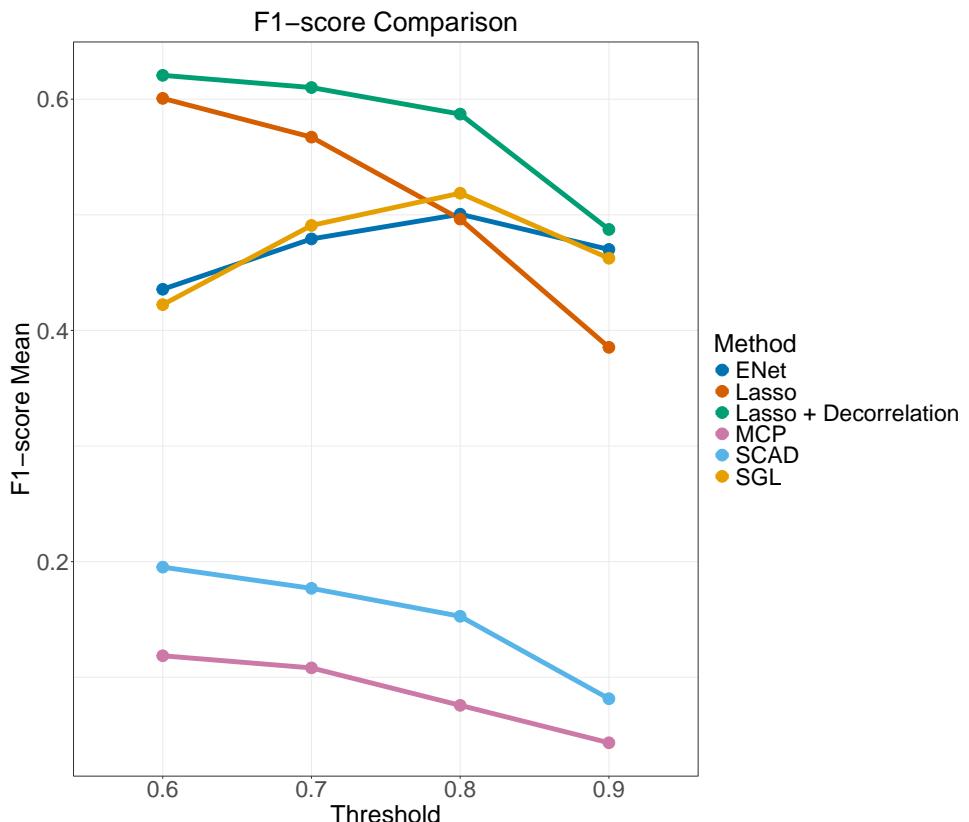
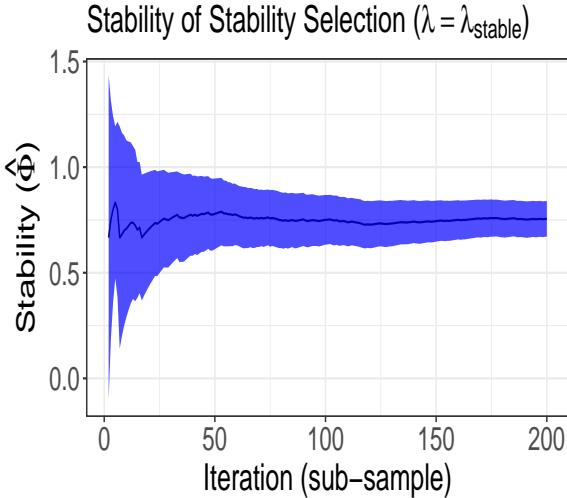
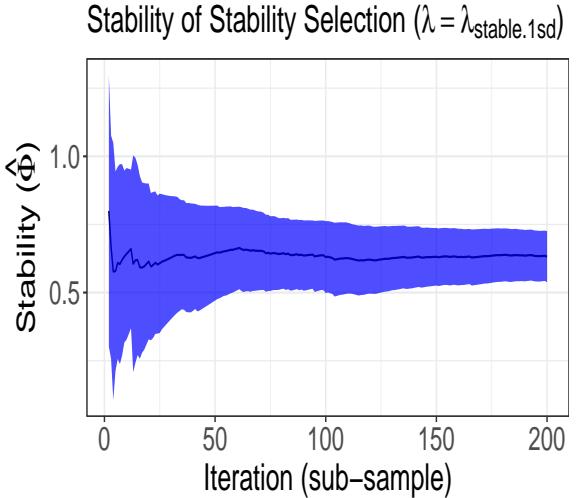


Figure 6: Comparison of F1-score across stability selection with different selection algorithms and decision-making thresholds



(a) Riboflavin data



(b) Microarray data

Figure 7: Selection stability of stability selection with the Lasso over sequential sub-sampling on the real data

## Why PCA Is Not Suitable in This Context?

We now discuss our preference for Gram–Schmidt process over Principal Component Analysis [PCA; T. and Jorge, 2016]. Although both methods produce orthogonal vectors, their underlying principles differ substantially. PCA operates on the covariance matrix of the data to identify directions that capture maximal variance, with each principal component formed as a linear combination of all original variables. In contrast, the Gram–Schmidt process orthogonalises a set of vectors by iteratively removing projections onto previously obtained vectors, thereby ensuring that each resulting vector remains directly associated with a specific original vector (variable). This property enables the use of the resulting orthonormal matrix  $Q$  as a proxy for performing variable selection on the original design matrix  $X$ . Therefore, although PCA is an effective tool for dimension reduction, it is not appropriate for the purpose of variable decorrelation in our context.

## Comparison with Forward Selection

The Gram–Schmidt process and forward stepwise selection are both iterative procedures, yet they are grounded in fundamentally different principles. The Gram–Schmidt process orthogonalises the variables, producing an orthonormal basis that removes correlation among variables. On the other hand, forward stepwise selection incrementally adds variables based on their correlation with the residuals, with the objective of maximising predictive performance. Notably, Gram–Schmidt identifies informative directions without reference to the response variable, whereas forward selection explicitly targets variables that enhance model fit. However, since we propose to order the variables according to their predictive power, the procedure bears some resemblance to forward selection. Nonetheless, it differs fundamentally in that it does not involve iterative model fitting on the response variable.

## Connection to Stable Learning

Stable learning has emerged as a rapidly growing research area, focused on identifying patterns that remain robust despite variations in the distribution of training and testing data (see, e.g., [Shen et al. \[2020\]](#), [Zhang et al. \[2021\]](#), and [Cui and Athey \[2022\]](#)). Given the challenges posed by correlated variables within this field [[Shen et al., 2020](#)], we contend that the methodology presented in this paper is closely related to this developing area of research, providing a promising direction for future investigation.

Furthermore, although in this paper the decorrelation procedure is applied to the entire dataset, it may be more appropriate to perform the variable decorrelation separately within each environment, particularly in settings where the data comprise samples from distinct environments, which is a common scenario in the context of stable learning [[Shen et al., 2020](#)].

## Connection to the Rashomon Sets

The index-sensitivity of the Gram–Schmidt process, which constructs an orthonormal set of variables based on their ordering, is reminiscent of Rashomon sets—collections of models that yield similarly high predictive performance using different subsets of variables [[Breiman, 2001](#)]. Although this topic lies beyond the scope of this paper, it would be of interest to investigate how different orderings of the design matrix lead to distinct predictive models, and how this phenomenon relates to the concept of Rashomon sets, for example in the context of [Kissel and Mentch \[2024\]](#).

## Future Work

Although in this paper the methodology is presented in the context of linear regression, it is evident that correlated variables also pose challenges in non-linear settings [[Gregorutti et al., 2017](#)]. Although addressing this issue is beyond the scope of this paper, we recommend further exploration in this area as a promising direction for future research. Furthermore, although the methodology has been presented within a regularisation framework, we hypothesise that the proposed approach could also be effective for non-regularisation-based methods, such as filter or wrapper methods [[Guyon and Elisseeff, 2003](#)], as correlated variables may introduce noise in these methods as well. This suggests an additional promising direction for future research.

## Competing interests

The authors declare that they have no conflict of interest.

## Author contributions statement

Mahdi Nouraie was responsible for drafting the manuscript, the development of the research methodology and for writing the computer code used throughout. Samuel Muller and Connor Smith provided critical feedback on the content of the manuscript, refining the clarity and scope of the manuscript and the computer code.

# Data Availability

The Riboflavin dataset is accessible via the `hdi` package in R [Dezeure et al., 2015]. The rat microarray data can be obtained from the National Center for Biotechnology Information (NCBI) website at [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov), under accession number GSE5680.

The source code used for the paper is accessible through the following GitHub repository: <https://github.com/MahdiNouraie/Variable-Decorrelation>. Furthermore, the `DVS` R package, which facilitates the use of the methodology introduced in this paper, is available through <https://github.com/MahdiNouraie/DVS>.

# Acknowledgments

Mahdi Nouraie was supported by the Macquarie University Research Excellence Scholarship (20213605). Samuel Muller was supported by the Australian Research Council Discovery Project Grant (DP230101908).

# References

- Å. Björck. Numerics of Gram-Schmidt Orthogonalization. *Linear Algebra and its Applications*, 197-198:297–316, 1994.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.
- Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014.
- Si-Bao Chen, Chris Ding, Bin Luo, and Ying Xie. Uncorrelated Lasso. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):166–172, 2013.
- Annie P Chiang, John S Beck, Hsan-Jan Yen, Marwan K Tayeh, Todd E Scheetz, Ruth E Swiderski, Darryl Y Nishimura, Terry A Braun, Kwang-Youn A Kim, Jian Huang, et al. Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292, 2006.
- Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4:110–115, 2022.
- Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statistical Science*, 30(4):533–558, 2015.
- Gregory Faletto and Jacob Bien. Cluster stability selection. *arXiv preprint arXiv:2201.00494*, 2022.

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

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 2013.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27:659–678, 2017.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618, 2008.

Ibrahim Joudah, Samuel Muller, and Houying Zhu. Air-HOLP: adaptive regularized feature screening for high dimensional correlated data. *Statistics and Computing*, 35(63), 2025.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12:95–116, 2007.

Nicholas Kissel and Lucas Mentch. Forward stability and model path selection. *Statistics and Computing*, 34, 2024.

Chenlei Leng, Yi Lin, and Grace Wahba. A note on the Lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

Ruiping Liu, Huiwen Wang, and Shanshan Wang and. Functional variable selection via Gram–Schmidt orthogonalization for multiple functional linear regression. *Journal of Statistical Computation and Simulation*, 88(18):3664–3680, 2018.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.

Mahdi Nouraie and Samuel Muller. On the Selection Stability of Stability Selection and Its Applications. *arXiv preprint arXiv:2411.09097*, 2024.

Vilfredo Pareto. *Cours d'Économie Politique*. Rouge, Lausanne, 1896.

Juliane Schafer, Rainer Opgen-Rhein, Verena Zuber, Miika Ahdesmaki, A. Pedro Duarte Silva, and Korbinian Strimmer. *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*, 2021. URL <https://CRAN.R-project.org/package=corpcor>. R package version 1.6.10.

Todd E Scheetz, Kwang-Youn A Kim, Ruth E Swiderski, Alisdair R Philp, Terry A Braun, Kevin L Knudtson, Anne M Dorrance, Gerald F DiBona, Jian Huang, Thomas L Casavant, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(1):55–80, 2013.

Zheyuan Shen, Peng Cui, Tong Zhang, and Kun Kunag. Stable learning via sample reweighting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5692–5699, 2020.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Joakim Skogholst, Kristian H. Liland, Tormod Næs, Age K. Smilde, and Ulf G. Indahl. Selection of principal variables through a modified Gram–Schmidt process with and without supervision. *Journal of Chemometrics*, 37(10):e3510, 2023.

Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging provides assumption-free stability. *Journal of Machine Learning Research*, 25(131):1–35, 2024.

Hervé Stoppiglia, Gérard Dreyfus, Rémi Dubois, and Yacine Oussar. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414, 2003.

Wei Sun, Junhui Wang, and Yixin Fang. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, 14(1):3419–3440, 2013.

Jolliffe Ian T. and Cadima Jorge. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(20150202), 2016.

Masaaki Takada, Taiji Suzuki, and Hironori Fujisawa. Independently Interpretable Lasso: A new regularizer for sparse regression with uncorrelated variables. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 454–463, 2018.

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

Haohan Wang, Benjamin J Lengerich, Bryon Aragam, and Eric P Xing. Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics*, 35(7):1181–1187, 2018.

Huiwen Wang, Ruiping Liu, Shanshan Wang, Zhichao Wang, and Gilbert Saporta. Ultra-high dimensional variable screening via Gram–Schmidt orthogonalization. *Computational Statistics*, 35(1153–1170), 2020.

Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):589–611, 2015.

Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:187–193, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

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

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyang Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, June 2021.

Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

## Appendix

### Gram-Schmidt Process

---

#### Algorithm 2 Gram-Schmidt Orthonormalisation

---

```

1: Input: A matrix  $X \in \text{mat}(n, p)$  with columns  $X_1, X_2, \dots, X_p$ 
2: Output: An orthonormal matrix  $Q \in \text{mat}(n, p)$  and an upper triangular matrix  $R \in \text{mat}(p, p)$ 
3: Initialise: Set  $Q$  and  $R$  as zero matrices.
4: for  $j = 1$  to  $p$  do
5:   Set  $v = X_j$ 
6:   if  $j > 1$  then
7:     for  $i = 1$  to  $j - 1$  do
8:       Compute  $R_{ij} = Q_i^\top X_j$ 
9:       Update  $v = v - R_{ij}Q_i$ 
10:    end for
11:   end if
12:   Compute  $R_{jj} = \|v\|_2$ 
13:   Normalise  $Q_j = \frac{v}{R_{jj}}$ 
14: end for

```

---

## Proof of Corollary 1

*Proof.* Here, we denote the ordered design matrix by  $X$ . Since the Gram-Schmidt process decomposes  $X$  as  $X = QR$ , by substituting  $Q$  as the new design matrix, the linear model  $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$  can be rewritten as  $\mathbf{Y} = Q(R\boldsymbol{\beta}) + \boldsymbol{\epsilon}$ . The irrepresentable condition for this form is  $\|Q_N^\top Q_S(Q_S^\top Q_S)^{-1}\|_\infty < 1$ . By the orthogonality of the columns of  $Q$  and the disjointness of the signal and noise variable sets, we have  $Q_N^\top Q_S = 0 \in \text{mat}(|N| \times |S|)$ , and  $Q_S^\top Q_S = \mathbb{I} \in \text{mat}(|S| \times |S|)$ , where  $|S|$  denotes the cardinality of the signal set  $S$ . Therefore, it follows that  $\|Q_N^\top Q_S(Q_S^\top Q_S)^{-1}\|_\infty = 0 < 1$ .  $\square$

## Proof of Corollary 2

*Proof.* By applying Algorithm 2 to  $X$  and keeping the matrix  $Q$ , we know that  $Q_1$  contributes more to the span of  $Q$  than  $Q_2$ . This is because  $X_2$  must provide some information that is not expressed by  $X_1$ . Since we assumed the two variables are equally relevant to the response variable, this implies that, from our linear modelling perspective, they are essentially the same. Therefore,  $Q_2$  is merely an orthogonal vector to  $Q_1$  and does not contribute to the prediction of the response variable  $\mathbf{Y}$ .

Given the orthogonality of  $Q$ , the Lasso solution treats each variable independently, with each coefficient determined solely by the inner product  $Q_j^\top Y$ . For  $Q_1$  and  $Q_3$ , these inner products equal  $\beta_1$  and 1, up to noise. Since both values exceed a reasonable regularisation value  $\lambda$ , the corresponding coefficients are retained in the solution. In contrast,  $Q_2$  is not associated with any signal, so  $Q_2^\top Y$  contains only noise and remains below the threshold for a reasonable regularisation value  $\lambda$ . As a result, the model consistently selects only  $Q_1$  and  $Q_3$ .

The variance estimator  $s_j^2$  is maximised when  $\overline{M(\lambda)_j} = 0.5$  and decreases as  $\overline{M(\lambda)_j}$  deviates from 0.5. The selection frequencies of the first two variables using  $X$  will converge to 0.5, as established in Corollary 2 of Faletto and Bien [2022]. In contrast, as demonstrated in the last paragraph, replacing  $X$  with  $Q$  results in a shift away from this maximum variance for the first two variables. Consequently,  $s_j^2$  becomes smaller for  $j = 1, 2$  when using  $Q$ . Given that the stability measure defined in Equation (2) is strictly decreasing in the total selection variance, the corollary is proved.  $\square$

## Beyond Lasso

Here, we evaluate our method on some other regularisation-based variable selection models. We use the second simulation scenario where in each group of correlated variables, the last index corresponds to an active variable. For all selection techniques, the regularisation value is determined using  $\lambda_{\text{stable-1sd}}$ .

### Elastic Net

We compare stability selection using the ENet with and without the variable decorrelation. The mixing parameter  $\text{alpha} = 0.2$ . Figure 8a highlights that employing our method significantly enhances the stability of selection results. In addition, Figure 8b illustrates that the F1-score

obtained with variable decorrelation surpasses that of the alternative method at lower threshold levels, but becomes inferior at higher thresholds.

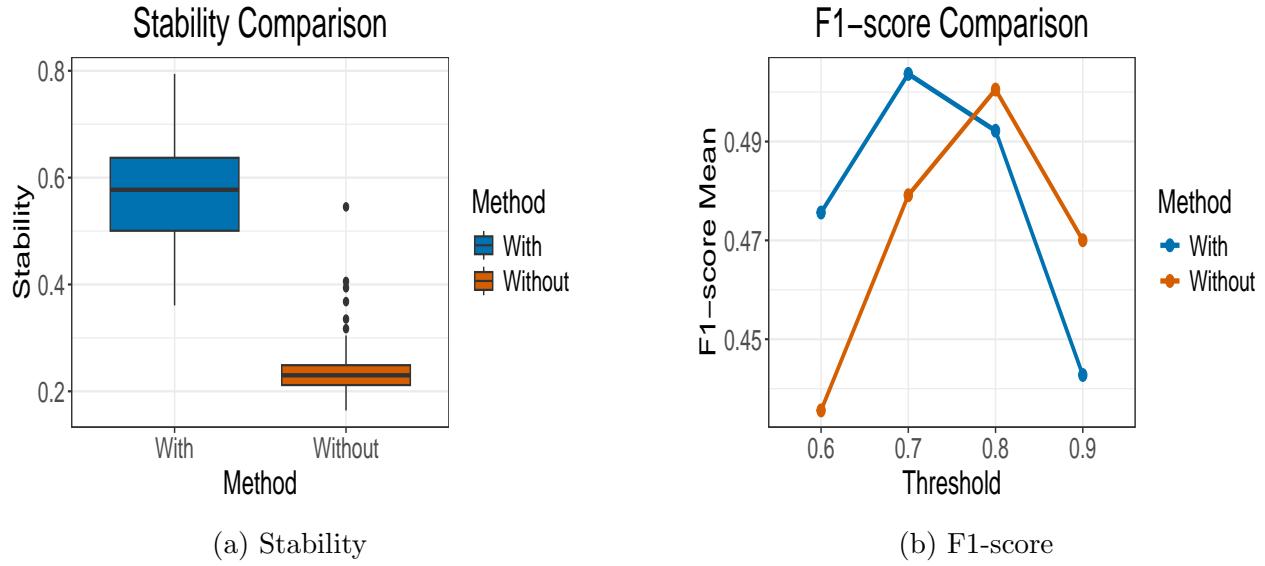


Figure 8: Comparing Stability and F1-score between stability selection using the ENet with and without the variable decorrelation for the second simulation scenario

### Smoothly Clipped Absolute Deviation

Figure 9 demonstrates that variable decorrelation leads to improvements in both stability and accuracy.

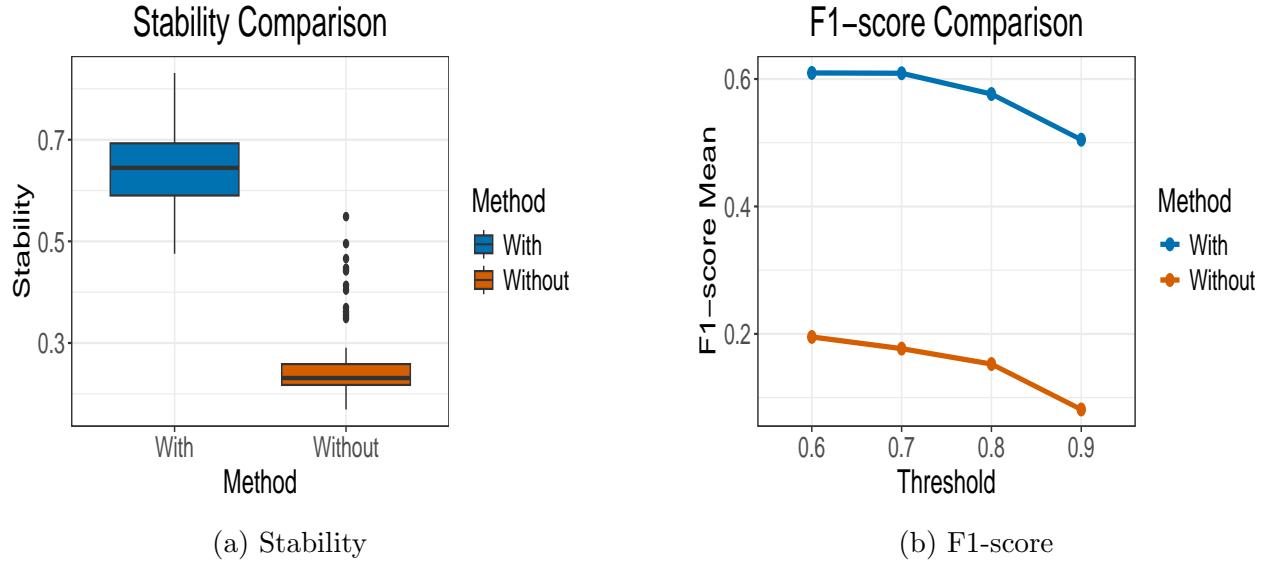


Figure 9: Comparing Stability and F1-score between stability selection using the SCAD with and without the variable decorrelation for the second simulation scenario

### Minimax Concave Penalty

Figure 10 presents patterns similar to Figure 9.

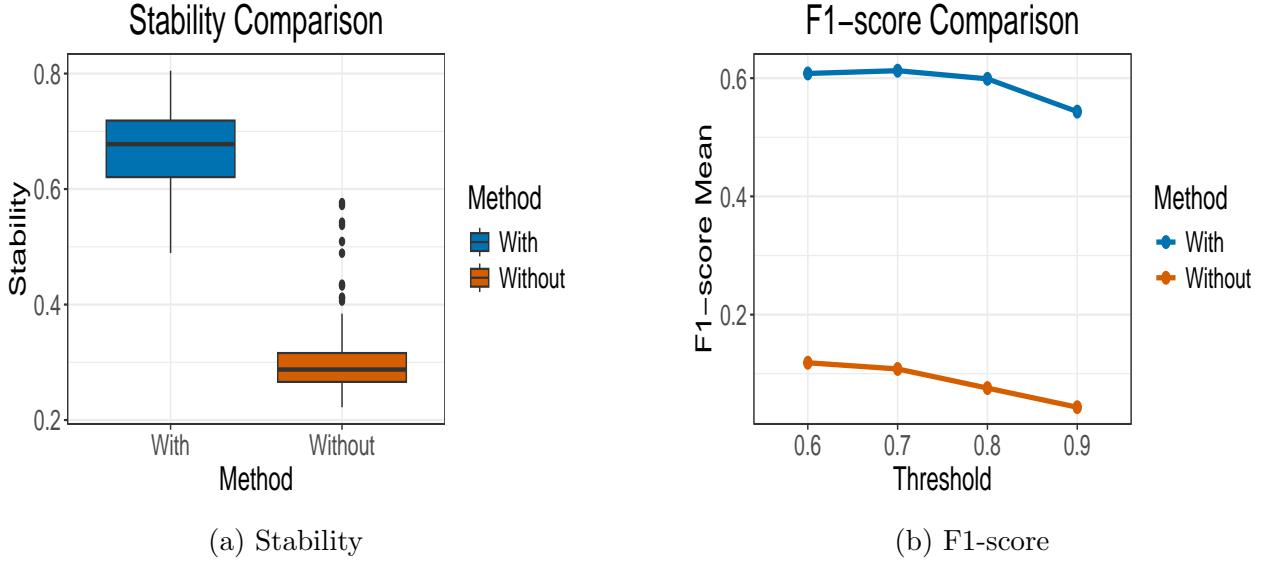


Figure 10: Comparing Stability and F1-score between stability selection using the MCP with and without the variable decorrelation for the second simulation scenario

### Sparse Group Lasso

The SGL encourages sparsity both between and within groups of correlated variables. Since the SGL requires the grouping structure of the variables, we provide the grouping index information to the model beforehand. Figure 11a illustrates that our method yields a slight positive effect on the stability of the results when compared to the previous selection techniques. This outcome is expected, as the loss function of the SGL is specifically designed to guide the model through group-related ambiguities by incorporating group indices and addressing sparsity at both the between- and within-group levels. However, it is noteworthy that our method still provides some advantages, even when applied to such a complex model. Figure 11b indicates that the selection F1-score improves when the variable decorrelation is applied and decision-making threshold is low. However, at higher thresholds, the decorrelation negatively affects the F1-score, similar to the pattern observed in Figure 8b.

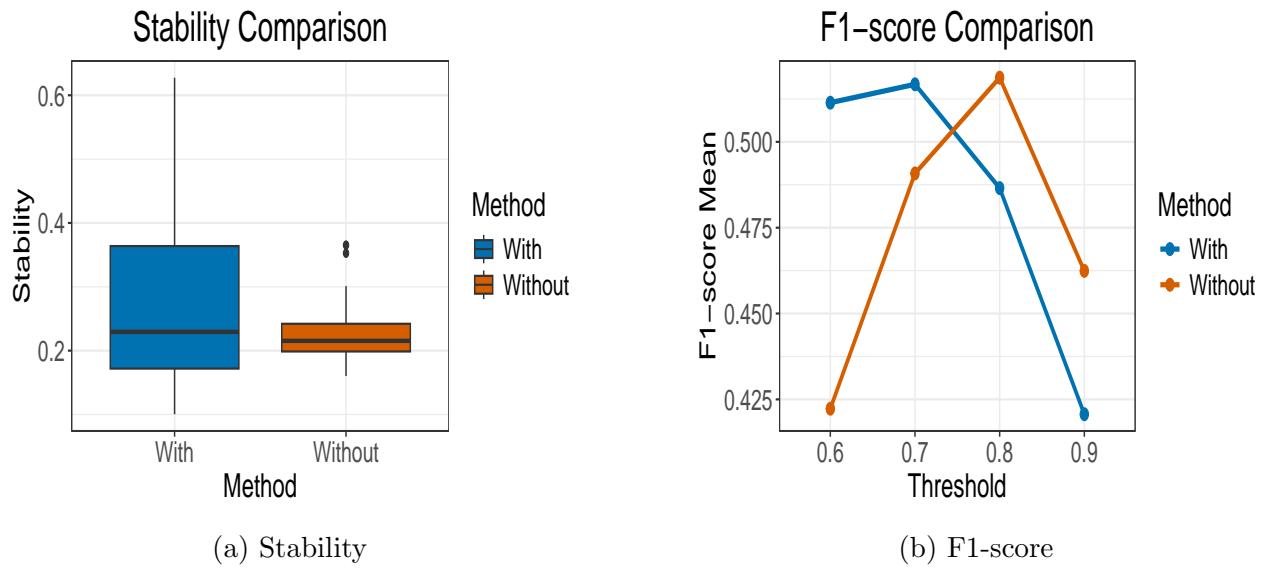


Figure 11: Comparing Stability and F1-score between stability selection using the SGL with and without the variable decorrelation for the second simulation scenario