VARIABLE SCREENING VIA TRUNCATED COMMON FACTORS FOR HIGH-DIMENSIONAL DATA WITH MULTICOLLINEARITY

Tanaka, Shuntaro
The Japan Research Institute, Limited.

Matsui, Hidetoshi
Faculty of Data Science, Shiga University

https://doi.org/10.5109/7238790

出版情報：Bulletin of informatics and cybernetics. 56 (6), pp.1-16, 2024. 統計科学研究会
バージョン：
権利関係：

VARIABLE SCREENING VIA TRUNCATED COMMON FACTORS FOR HIGH-DIMENSIONAL DATA WITH MULTICOLLINEARITY

by

Shuntaro Tanaka and **Hidetoshi Matsui**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.56, No. 6

FUKUOKA, JAPAN
2024

Bulletin of Informatics and Cybernetics, Vol. 56, No. 6, 2024

VARIABLE SCREENING VIA TRUNCATED COMMON FACTORS FOR HIGH-DIMENSIONAL DATA WITH MULTICOLLINEARITY

By

Shuntaro Tanaka[*†] and **Hidetoshi Matsui**[‡]

Abstract

Screening methods are useful tools for variable selection in regression analysis when the number of predictors is much larger than the sample size. Factor analysis is used to eliminate multicollinearity among predictors, which improves the variable selection performance. We propose a new method, called Truncated Preconditioned Profiled Independence Screening that better selects the number of factors to eliminate multicollinearity. The proposed method improves the variable selection performance by truncating unnecessary parts from the information obtained by factor analysis. We confirmed the superior performance of the proposed method in variable selection through analysis using simulation data and real datasets.

Key Words and Phrases: Factor analysis, High dimensional data, Multicollinearity, Screening, Variable selection.

1. Introduction

Recent developments in the field of communication technology have generated data in a variety of fields, including finance, medicine, and agriculture. Appropriate analysis of such data enables us to reveal the relationships inherent in the complex phenomena. Regression analysis is one of the most widely used statistical methods to do this (Hastie et al., 2009; Fahrmeir et al., 2013). For example, if we want to understand the regularity of the sales of a product, we set the sales as the response and the product attributes (price, color, size, etc.) as the predictors. To understand the correct relationship between the predictors and the response, it is necessary to select and analyze important variables from the large amount of data that appear to be strongly related to a given response.

Variable selection is used in several fields. In finance, variables related to corporate accounting data are selected to construct a statistical model that predicts the risk of corporate bankruptcy (Tian et al., 2015). Another example is the selection of variables that relate to data on macroeconomic indicators to estimate volatility, which is used to select which company to invest in and to make decisions about the timing of investments

[*] The Japan Research Institute, Ltd. 2–18–1 Higashi–gotanda Shinagawa–ku Tokyo Japan. tanaka.shuntaro@jri.co.jp

[†] Graduate School of Data Science, Shiga University 1–1–1 Banba Hikone Shiga 522–8522 Japan. s7022103@st.shiga-u.ac.jp

[‡] Faculty of Data Science, Shiga University 1–1–1 Banba Hikone Shiga 522–8522 Japan. hmatsui@biwako.shiga-u.ac.jp

(Fang et al., 2020). Variable selection is also used in clinical models that predict possible future diseases (Chowdhury and Turin, 2020) and in near-infrared spectroscopy analysis to measure food compositions (Yun et al., 2019).

It is difficult to apply the classical variable selection techniques such as stepwise regression to high-dimensional data. LASSO (Tibshirani, 1996), the typical example of methods using L_1-type regularization, also has an issue that when the number of variables p exceeds the sample size n, it selects only n variables at most. More recently, Sure Independence Screening (SIS) was proposed to greatly reduce the dimension of the predictors and select important variables (Fan and Lv, 2008). SIS selects predictors in the order of their Pearson's correlations with the response in linear regression models. Although this is a simple technique, the probability that the set of variables selected by SIS contains a set of truly important variables converges to one as the sample size increases. Several extensions of SIS have been proposed. Fan and Song (2010) extended the idea of SIS to generalized linear models, and Fan et al. (2011) extended it to high-dimensional additive models. In addition, there are screening methods that use nonlinear correlations instead of Pearson correlations. Li et al. (2012a) proposed a method that is robust to outliers that uses Kendall's rank correlation coefficient. Li et al. (2012b) used distance correlation, and Balasubramanian et al. (2013) used the Hilbert-Schmidt Independence Criterion (HSIC). With these criteria, we can apply the screening methods without assuming any distribution for the variables. Zhang et al. (2017) also proposed a method for censored data. The development of screening methods was summarized in Fan and Lv (2018).

Nevertheless, most of these screening methods have the problem that their performance degrades in the presence of multicollinearity. To solve this problem, Wang and Leng (2016) proposed a method called High-dimensional Ordinary Least squares Projection (HOLP), which accommodates highly multicollinear predictors by selecting variables in the order of their relations estimated by high-dimensional ordinary least squares. Factor Profiled Sure Independence Screening (FPSIS) proposed by Wang (2012) transforms the data for predictors by applying factor analysis, which reduces multicollinearity. Then we can select appropriate variables by applying SIS to the transformed data that correspond to unique factors. Preconditioned Profiled Independence Screening (PPIS) proposed by Zhao et al. (2020) improved the FPSIS transformation process to reduce multicollinearity more correctly. PPIS eliminates unnecessary information from the predictors by using all of the common factors obtained from applying factor analysis to the predictors, whereas FPSIS uses only a subset of common factors.

However, PPIS seems to eliminate more information about predictors than necessary, which can degrade variable selection performance. To overcome this issue, we propose a method to improve the effectiveness of removing multicollinearity by modifying PPIS to select variables more accurately. We truncate some of the common factors eliminated in the PPIS transformation process to prevent excessive loss of information for variable screening. We call our proposed method Truncated PPIS (TPPIS). The reason why TPPIS improves the variable selection performance can be explained by a model based on the distribution of eigenvalues. The truncation part is determined objectively using the BIC-type criterion proposed by Wang (2012). SIS is then applied to the data whose multicollinearity has been removed by the transformation process. Through analysis of simulated and real data, we show that TPPIS can transform data appropriately.

The remainder of this paper is organized as follows. Section 2 describes existing screening methods, and then the proposed method is described in Section 3. In Section 4, we confirm the performance of the screening method through a simulated data analysis, and then report the results of real data analysis in Section 5. Section 6 presents discussions.

2. Screening methods utilizing factor analysis

Suppose we have n sets of observations $\{(y_i, \boldsymbol{x}_i);\ i = 1, \ldots, n\}$, where $y_i \in \mathbb{R}$ is a response and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is a vector of predictors. In particular, we assume that $n < p$ and $\boldsymbol{x}_i$ is standardized and y_i is centered. The relationship between y_i and $\boldsymbol{x}_i$ is assumed to be represented by the following linear model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ are regression coefficients and $\varepsilon_i \in \mathbb{R}$ is independent and identically distributed (i.i.d.) random noise following $N(0, \sigma^2)$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. Then the above linear model can be expressed as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1}$$

Fan and Lv (2008) defined the set of indices of truly important variables as $M^* = \{1 \le j \le p : \beta_j \neq 0\}$. Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^\top = X^\top \boldsymbol{y} \in \mathbb{R}^p$ and define the importance of the j-th variable as $|\omega_j|$ $(1 \le j \le p)$. SIS excludes predictors that are considered to be unnecessary by selecting the j-th variables in order of increasing $|\omega_j|$. However, SIS does not work well in the presence of strong multicollinearity (Fan and Lv, 2008). For example, $|\omega_j|$ becomes smaller even for important variables, or $|\omega_j|$ becomes larger even for unimportant variables.

In FPSIS (Wang, 2012), SIS is applied after a transformation process to remove multicollinearity by applying factor analysis. Let $Z \in \mathbb{R}^{n \times d}$ be a matrix of vectors of d $(< n)$ common factors of X, $B \in \mathbb{R}^{p \times d}$ be factor loadings, and $\check{X} \in \mathbb{R}^{n \times p}$ be a matrix composed of unique factors. Then we can express their relationships as $X = ZB^\top + \check{X}$, where the columns of $\check{X}$ are independent of each other. Although Z is not uniquely determined due to the rotation invariance, a solution for Z can be obtained by singular value decomposition.

Let $\mu_1, \ldots, \mu_n$ be n singular values of X, where $\mu_1 \ge \ldots \ge \mu_n > 0$, since we assume $n < p$ here. The singular value decomposition of X gives

$$X = UDV^\top, \tag{2}$$

where $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \in \mathbb{R}^{n \times n}, \boldsymbol{u}_l = (u_{1l}, \ldots, u_{nl})^\top \in \mathbb{R}^n, D = \mathrm{diag}(\mu_1, \ldots, \mu_n) \in \mathbb{R}^{n \times n}, V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \in \mathbb{R}^{p \times n}, \boldsymbol{v}_l = (v_{1l}, \ldots, v_{pl})^\top \in \mathbb{R}^p$ $(l = 1, \ldots, n)$, and $U^\top U = V^\top V = I_n$. Let $U_1 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d) \in \mathbb{R}^{n \times d}$ denote the first d columns of the matrix U in (2). Then U_1 can be regarded as one of the solutions of Z. Wang (2012) decided the value of d by the following equation using the ratio of the singular values of X:

$$d = \underset{1 \le l \le n-1}{\mathrm{argmax}} \frac{\mu_l^2}{\mu_{l+1}^2}. \tag{3}$$

The projection matrix onto the orthogonal complement of the linear subspace spanned by the column vectors of the matrix U_1 is given by

$$Q_F = I_n - U_1 U_1^\top. \tag{4}$$

Left-multiplying both sides of (1) by Q_F gives

$$Q_F \boldsymbol{y} = Q_F X \boldsymbol{\beta} + Q_F \boldsymbol{\varepsilon}. \tag{5}$$

Let $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\top = Q_F \boldsymbol{y}$ and $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_p) = Q_F X$. $\hat{X}$ is an approximation of the unique factors $\check{X}$. The use of $\hat{X}$ instead of X enables us to eliminate multicollinearity and to select appropriate variables. FPSIS calculates $\omega_j = (n^{-1}\hat{X}_j^\top \hat{X}_j)^{-1}(n^{-1}\hat{X}_j^\top \hat{\boldsymbol{y}}) = \hat{X}_j^\top \hat{\boldsymbol{y}} / \hat{X}_j^\top \hat{X}_j$, and then selects variables where $|\omega_j|$ is large in order.

PPIS (Zhao et al., 2020) improved the FPSIS transformation process. First, after applying SVD to X as in (2), they divided each of the matrices U, D, V into two parts at the d-th column: $U_1 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d) \in \mathbb{R}^{n \times d}$, $U_2 = (\boldsymbol{u}_{d+1}, \ldots \boldsymbol{u}_n) \in \mathbb{R}^{n \times (n-d)}$, $D_1 = \mathrm{diag}(\mu_1, \ldots, \mu_d) \in \mathbb{R}^{d \times d}$, $D_2 = \mathrm{diag}(\mu_{d+1}, \ldots, \mu_n) \in \mathbb{R}^{(n-d) \times (n-d)}$, $V_1 = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d) \in \mathbb{R}^{p \times d}$, $V_2 = (\boldsymbol{v}_{d+1}, \ldots \boldsymbol{v}_n) \in \mathbb{R}^{p \times (n-d)}$. Let

$$Q_P = U_2 D_2^{-1} U_2^\top \left(I_n - U_1 U_1^\top\right) = U_2 D_2^{-1} U_2^\top \tag{6}$$

and replace Q_F with Q_P in (5). This is based on the Puffer transformation (Jia and Rohe, 2012). PPIS calculates $\omega_j = \hat{X}_j^\top \hat{\boldsymbol{y}} / \hat{X}_j^\top \hat{X}_j$ as in FPSIS, where $\hat{\boldsymbol{y}} = Q_P \boldsymbol{y}$, $\hat{X} = Q_P X$, and then selects variables in order of the size of $|\omega_j|$. The number of dimensions d of U_1 is determined by (3) using the ratio of the singular values of X. We explain the reasonableness of PPIS in Section 3.2. using a model based on the distribution of eigenvalues.

However, if the magnitudes of the singular values after the d-th are not sufficiently small compared to those before the d-th, $\hat{X}$ is still multicollinear when we simply remove from X the effects that are related to the first d common factors of X. Therefore, by removing the influence of the n common factors of X including the information after the d-th factor that is not used in FPSIS, $\hat{X}$ becomes closer to the unique factors $\check{X}$, which leads to the elimination of more multicollinearity.

3. Proposed method

3.1. TPPIS

We propose selecting the number of factors to eliminate more multicollinearity by modifying the transformation process in PPIS. Let α be a tuning parameter that satisfies $\alpha \in (0, 1]$ and $d < [n\alpha]$. After applying SVD to X, as in (2), we divide U, D, V into three parts at the d-th column and the $[n\alpha]$-th column: $U = (U_1, U_{2a}, U_{2b})$, $D = \mathrm{blockdiag}\{D_1, D_{2a}, D_{2b}\}$, $V = (V_1, V_{2a}, V_{2b})$, $U_1 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)$, $U_{2a} = (\boldsymbol{u}_{d+1}, \ldots, \boldsymbol{u}_{[n\alpha]})$, $U_{2b} = (\boldsymbol{u}_{[n\alpha]+1}, \ldots, \boldsymbol{u}_n)$, $D_1 = \mathrm{diag}(\mu_1, \ldots, \mu_d)$, $D_{2a} = \mathrm{diag}(\mu_{d+1}, \ldots, \mu_{[n\alpha]})$, $D_{2b} = \mathrm{diag}(\mu_{[n\alpha]+1}, \ldots, \mu_n)$, $V_1 = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$, $V_{2a} = (\boldsymbol{v}_{d+1}, \ldots, \boldsymbol{v}_{[n\alpha]})$, and $V_{2b} = (\boldsymbol{v}_{[n\alpha]+1}, \ldots, \boldsymbol{v}_n)$. Then we define the following projection matrix

$$Q_T = U_{2a} D_{2a}^{-1} U_{2a}^\top \left(I_n - U_1 U_1^\top\right) = U_{2a} D_{2a}^{-1} U_{2a}^\top.$$

Using $\hat{X} = Q_T X$ rather than $Q_P X$, we can eliminate multicollinearity more accurately since Q_T leaves the information that corresponds to the unique factors by truncating

U_{2b} and D_{2b} from U_2 and D_2, respectively. TPPIS calculates $\hat{\boldsymbol{y}}, \hat{X}$, and $\boldsymbol{\omega}$ using the equation that replaces Q_F with Q_T in (5), and then selects variables where $|\omega_j|$ is large in order to identify the set of indices of important variables that satisfy $\beta_j \neq 0$ in (1).

Denote a set of k selected variables as

$$M_k = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } k \text{ largest of all } \}$$

and denote predictors whose columns are composed of M_k as $X(M_k) \in \mathbb{R}^{n \times k}$. We predict the response using $\boldsymbol{y} = X(M_k)\hat{\boldsymbol{\beta}}(M_k)$, where $\hat{\boldsymbol{\beta}}(M_k)$ is the least squares estimator of the regression coefficient of $X(M_k)$, that is,

$$\hat{\boldsymbol{\beta}}(M_k) = \left\{X(M_k)^\top X(M_k)\right\}^{-1} X(M_k)^\top \boldsymbol{y}. \tag{7}$$

3.2. Reasons why TPPIS improves the effectiveness of removing multicollinearity

We discuss the reason why TPPIS improves the effectiveness of removing multicollinearity and the variable selection performance. Zhao et al. (2020) indicates that the transformation process using Q_P of (6) works well for data that follow a highly multicollinear spike model. The spike model has the property that some eigenvalues of the variance-covariance matrix are larger than others. Suppose that the eigenvalues of a variance-covariance matrix X, denoted by Σ_p, can be divided into three size categories: large, medium, and small. Among p eigenvalues, let d be the number of large eigenvalues, m be the number of medium eigenvalues, and $p-d-m$ be the number of small eigenvalues. Then the spike model assumes that Σ_p is represented as

$$\Sigma_p = \sum_{r=1}^{d}(\lambda_r + \sigma_0^2)\boldsymbol{u}_r^*\boldsymbol{u}_r^{*\top} + \sum_{s=1}^{m}(\omega_s + \sigma_0^2)\boldsymbol{u}_{d+s}^*\boldsymbol{u}^{*\top}_{d+s} + \sum_{t=1}^{p-d-m}\sigma_0^2\boldsymbol{u}_{d+m+t}^*\boldsymbol{u}^{*\top}_{d+m+t},$$

where $\lambda_1 \geq \ldots \geq \lambda_d > \omega_1 \geq \ldots \geq \omega_m > 0$, σ_0^2 is a positive constant, and $\{\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_p^*\}$ constitute an orthonormal basis of $\mathbb{R}^p$. In this case, X can be expressed as

$$X = \sum_{r=1}^{d}\sqrt{\lambda_r}\boldsymbol{z}_r\boldsymbol{u}_r^{*\top} + \sum_{s=1}^{m}\sqrt{\omega_s}\boldsymbol{z}_{d+s}\boldsymbol{u}_{d+s}^{*\top} + \sigma_0^2\Lambda, \tag{8}$$

where $\boldsymbol{z}_w \in \mathbb{R}^n$ $(w = 1, \ldots, d+m)$ are i.i.d. $N(\boldsymbol{0}, I_n)$ vectors and $\Lambda \in \mathbb{R}^{n \times p}$ has i.i.d. $N(0,1)$ elements. The vectors $\boldsymbol{z}_r$ and $\boldsymbol{u}_r^*$ respectively represent a common factor and a factor loading of X, and $\sigma_0^2\Lambda$ represents a unique factor of X. Let X_1, X_2, X_3 be the first, second, and third terms of (8), respectively; that is, we can express (8) as $X = X_1 + X_2 + X_3$.

Since Q_F in (4) is the projection matrix onto the orthogonal complement of the linear subspace spanned by the column vector $U_1 \in \mathbb{R}^{n \times d}$, Q_F can remove the effect of d common factors. That is,

$$\begin{aligned} Q_F X &= Q_F(X_1 + X_2 + X_3) \\ &\approx X_2 + X_3. \end{aligned}$$

Similar to Q_F, the transformation by the matrix $\left(I_n - U_1U_1^\top\right)$ in Q_P in (6) eliminates the components relating to X_1 from X. The matrix $U_2D_2^{-1}U_2^\top$ in Q_P, based on the

Puffer transformation, makes singular values of X close to each other, which leads to the elimination of correlations between the columns. Therefore, Q_P can eliminate the components relating to X_2 and X_3. Since X_3 is associated with unique factors, it should be retained as much as possible. However, since U_2 and D_2 utilize all the information beyond the d-th dimension, Q_P unnecessarily removes additional components relating to X_3. In contrast, Q_T preserves the components of X_3 by truncating U_{2b} and D_{2b}, then TPPIS effectively eliminate the multicollinearity.

3.3. Selection of tuning parameters

The performance of the proposed method strongly depends on the dimension d of U_1, the tuning parameter α, and the number k of selected variables. We have to decide appropriate values for them. To do this, we use the BIC-type criterion adapted to high-dimensional data proposed by Wang (2012). Using $\hat{\boldsymbol{\beta}}(M_k)$ in (7), the BIC-type criterion is given by

$$\mathrm{BIC}(M_k) = \log\left\{\left|\left|\boldsymbol{y} - X(M_k)\hat{\boldsymbol{\beta}}(M_k)\right|\right|^2\right\} + (n^{-1}\log p)|M_k|\ \log n. \tag{9}$$

We use grid search to find the optimal d, α, and k, selecting the values with which make BIC smallest as the optimal parameters.

4. Simulation examples

To investigate the effectiveness of the proposed TPPIS method, we compare TPPIS with the existing methods. After calculating the importance of each predictor on the response for each method, the number of variables is determined using the BIC-type criterion (9), and then the variable selection performance is verified.

4.1. Settings for simulated data

Following the simulations in Zhao et al. (2020), we performed simulation studies in four settings. The sample size n and the number of predictors p are set as $n = 100, 300$, and $p = 1000$ as common values for each example, respectively. For the TPPIS parameter d, we examined five patterns: $0.2n, 0.4n, 0.6n, 0.8n$, and the value given by (3). In addition, we examined five values ranging from 0.2 to 1.0 in increments of 0.2 for α. In the numerical experiments, we consider only the values of d and α such that $d < [n\alpha]$. For the number of variables, k, we examined n values ranging from 1 to n. We then select the d, α, and k giving the smallest BIC as the optimal parameters.

- Example 1

  In this case, among the 1000 predictors, four variables are related to the response. For each i in $1 \leq i \leq n$,

  $$y_i = 5x_{i1} + 5x_{i2} + 5x_{i3} - 15x_{i4} + \varepsilon_i,$$

  where ε_i are i.i.d. errors following $N(0, 1)$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ are i.i.d. predictors

following $N(\mathbf{0}, \Sigma)$ and the variance-covariance matrix $\Sigma = (\Sigma_{jk})_{j,k=1}^{p}$ satisfies

$$\begin{aligned}\Sigma_{jj} &= 1,\\ \Sigma_{jk} &= \varphi \ (j \neq k, j \neq 4, k \neq 4),\\ \Sigma_{4,k} &= \Sigma_{j,4} = \sqrt{\varphi} \ (j, k \neq 4).\end{aligned}$$

We investigated three values for the parameter φ: 0.5, 0.7, and 0.9.

- Example 2

  For each i in $1 \leq i \leq n$,

  $$y_i = 5x_{i1} + 5x_{i2} + 5x_{i3} - 15x_{i4} + 5x_{i5} + \varepsilon_i.$$

  The setting is similar to that in Example 1, but the fifth variable is added. In addition, the variance-covariance matrix Σ of the predictor satisfies $\Sigma_{5,j} = \Sigma_{j,5} = 0$ $(j \neq 5)$.

- Example 3

  For each i in $1 \leq i \leq n$,

  $$y_i = 5x_{i1} + 5x_{i2} + 5x_{i3} - 15x_{i4} + 5x_{i5} + \varepsilon_i.$$

  The regression model is the same as in Example 2, except that the sixth variable, which is not included in the regression model, satisfies $x_{i6} = 0.8x_{i5} + \delta_i$, where δ_i follows i.i.d. $N(0, 0.01)$. Compared to Example 2, the data for the predictors are more multicollinear.

- Example 4

  We consider the case where X follows a spike model (8), given by

  $$X = \sum_{r=1}^{d} \boldsymbol{z}_r \boldsymbol{b}_r^\top + \sum_{s=1}^{m} n^{\frac{-(s+9)}{m+10}} \boldsymbol{z}_{d+s} \boldsymbol{b}_{d+s}^\top + \check{X},$$

  where $\boldsymbol{z}_k \in \mathbb{R}^n$ $(k = 1, \ldots, d+m)$ are i.i.d. vectors following $N(\mathbf{0}, I_n)$, $\boldsymbol{b}_k \in \mathbb{R}^p$ is a vector of i.i.d. $N(0,1)$ elements, and $\check{X} = (\check{\boldsymbol{x}}_1, \ldots, \check{\boldsymbol{x}}_n)^\top \in \mathbb{R}^{n \times p}$ with $\check{\boldsymbol{x}}_i = (\check{x}_{i1}, \ldots, \check{x}_{ip})^\top \in \mathbb{R}^p$, $E(\check{x}_{ij}) = 0$, and $\mathrm{cov}(\check{x}_{ij_1}, \check{x}_{ij_2}) = I_p$. This case corresponds to equation (8) with $\sqrt{\lambda_r} = 1$ $(1 \leq r \leq d)$, $\sqrt{\omega_s} = n^{\frac{-(s+9)}{m+10}}$ $(1 \leq s \leq m)$, and $\sigma_0^2 = 1$.

  In this example, d is set to 3 and m is set according to four patterns: $0.2n$, $0.4n$, $0.6n$, and $0.8n$. The regression model is given by

  $$y_i = 5x_{i1} + 4x_{i2} + 3x_{i3} + 2x_{i4} + \varepsilon_i,$$

  where ε_i are i.i.d. errors following $N(0, \sigma^2)$ with $\sigma^2 = \mathrm{var}(X\boldsymbol{\beta})/5$ and $\boldsymbol{\beta} = (5, 4, 3, 2, 0, \ldots, 0)^\top \in \mathbb{R}^p$.

In each example, we generate datasets 100 times for each combination of parameters. For each dataset, the numbers of selected predictors and the least squares estimator (7) is calculated. The number of variables is determined using the BIC in (9).

4.2. Comparison methods

The proposed TPPIS method is compared with the existing SIS, FPSIS, and PPIS methods. TPPIS involves two tuning parameters, d and α, to determine the transformation matrix Q_T. We examine two types of TPPIS: one is that d is determined by (3) and α is selected based on BIC in (9), and another is that both d and α are determined using BIC in (9). We refer to the former as TPPIS$_\alpha$ and the latter as TPPIS$_{d,\alpha}$. In addition to the original FPSIS, which selects the value of d by (3), we also compare a modified FPSIS where d is selected by the BIC in (9). We denote this method as FPSIS$_d$. We test the values of d in FPSIS$_d$ with five patterns, as in the case of TPPIS.

4.3. Score metric for screening

We evaluate the variable selection performance of the screening methods using the score based on the number of correctly and incorrectly selected variables. We refer to necessary predictors as Positive (P) and unnecessary variables as Negative (N) in the regression model. Since the true regression coefficients of the simulated data are known, we can calculate True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), Recall (TP/(TP+FN)), and Precision (TP/(TP+FP)).

The weighted F-score is weighted on the Recall side by the importance θ as follows:

$$F\theta\text{-score} = \frac{1+\theta^2}{\frac{1}{\text{Precision}} + \frac{\theta^2}{\text{Recall}}},$$

where Precision = TP/(TP+FP) and Recall = TP/(TP+FN). Since the screening methods aim to select variables while retaining as many necessary variables as possible, we focus on Recall and therefore we use F2-score.

4.4. Simulation results

The results of the variable selection for Example 1 are shown in Table 1. The numbers in the $x_{(j)}$ column represent the total number of times that the j-th predictor variable is selected. For all settings, SIS never selected $x_{(4)}$. This is because $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $(x_{14}, \ldots, x_{n4})^\top$ are uncorrelated due to the generation mechanism of the data, which gives a smaller $|\omega_4|$. For other methods than SIS, the value of $|\omega_4|$ is larger than that for SIS due to the transformation process by factor analysis. In particular, TPPIS$_\alpha$ selected $x_{(4)}$ more frequently than the existing methods, and moreover, TPPIS$_{d,\alpha}$ obtained the largest $x_{(4)}$. The F2-scores of TPPIS$_\alpha$ are better than those of existing methods, and those of TPPIS$_{d,\alpha}$ are the highest values in all settings. We confirmed that the performance of TPPIS in variable selection is improved compared to the existing methods. Figure 1 shows values of BIC and F2-scores for fixed d and different α in TPPIS. This figure demonstrates that α is selected appropriately by BIC.

The results for Example 2 are shown in Table 2. The table shows that in many cases the numbers in $x_{(5)}$ are close to 100 because the fifth variable is uncorrelated with the other predictors. In all cases, the F2-scores of TPPIS$_\alpha$ and TPPIS$_{d,\alpha}$ are the same and the highest.

Table 3 summarizes the result for Example 3. This shows that the numbers in $x_{(5)}$ and the value of F2-score are smaller than those of Example 2 due to the addition of the sixth variable, which is highly correlated with the fifth variable. For the cases with

$n = 300$ and $\varphi = 0.9$, FPSIS_d and $\text{TPPIS}_{d,\alpha}$, which determine d by BIC, give lower $x_{(6)}$ values. It seems to be useful to use BIC to select d for data with multicollinearity. $\text{TPPIS}_{d,\alpha}$ gives the highest F2-score among all methods.

Table 4 shows the results for Example 4. In this example, the variables with large regression coefficients tend to be more important, resulting in $x_{(1)} \geq x_{(2)} \geq x_{(3)} \geq x_{(4)}$ under many settings. F2-scores for PPIS and TPPIS are high because these methods are effective for the spike model. In particular, TPPIS_α and $\text{TPPIS}_{d,\alpha}$ give the highest F2-scores for all settings.

5. Real data analysis

We apply the proposed screening methods to the analysis of two real datasets. For both datasets, we selected variables by TPPIS_α and $\text{TPPIS}_{d,\alpha}$, where the tuning parameters d and α are determined by BIC from the same set of d and α candidates as in Section 4.1.

5.1. Condition monitoring of hydraulic systems

We applied the screening methods to data on condition monitoring of a hydraulic system (Helwig et al., 2015). This dataset was obtained experimentally using a hydraulic test rig to measure values such as pressure, volumetric flow, and temperature while varying the settings of four different hydraulic components (coolers, valves, pumps, and accumulators). We use data with the sample size 1449, taken under stable system settings. The response is a value that expresses the degree of accumulator failure as a continuous value. A higher value is closer to normal condition with 130 being the optimal pressure, 115 being a slightly reduced pressure, 100 being a severely reduced pressure, and 90 being close to total failure. The predictors are the values measured by 17 sensors and form a total of 43680. We apply the five screening methods to analyze this dataset as in the section on examples of simulated data. The number of variables is determined using BIC.

Table 5 shows the results of the analysis of this dataset. From this result, we find that TPPIS_α and $\text{TPPIS}_{d,\alpha}$ select variables from the largest number of sensors. These methods select variables 'volume flow sensors (FS)' and 'efficiency factor (SE),' which are not selected by the other methods. In addition, these methods give the best BIC score among all methods. These results indicate that these sensors may relate to the condition of accumulators.

5.2. S&P500

The S&P 500, one of the U.S. stock market indices, is obtained by weighting the market capitalization of 500 companies selected as representative of publicly traded companies. This analysis uses the data for the year 2020. The sample size is 253, which is the number of trading days. The response is the value of the S&P500, and the predictors are the stock price of each of the 500 companies that make up the S&P500. Note that the number of columns of predictors may be greater than 500 because some companies have multiple stocks, differentiated based on whether they include voting rights. Since the S&P500 is weighted by market capitalization, it is assumed that the stock price of the company with the highest market capitalization is selected as an important variable.

The values of the S&P500 are obtained from Federal Reserve Economic Data (FRED)[1], and the stock prices of the 500 companies that make up the S&P500 are obtained from the website[2].

We applied six screening methods to this dataset and compared BIC and selected variables. The results for the S&P500 are shown in Table 6. $\text{TPPIS}_{d,\alpha}$ gives the best BIC score among all the methods. The seven variables selected by $\text{TPPIS}_{d,\alpha}$ include companies with particularly large market capitalizations such as 'AAPL' (Apple), 'MSFT' (Microsoft) and 'AMZN' (Amazon).

6. Discussion

We have proposed TPPIS, a variable screening method for high-dimensional data with strong multicollinearity. TPPIS improves the variable selection performance by using a BIC-type criterion to determine the number of common factors that have a role in removing multicollinearity. In the analysis of simulated data, TPPIS outperformed existing methods using factor analysis for variable selection. This suggests that TPPIS may be able to correctly select variables that are not considered important by existing methods.

The transformation process of TPPIS to remove multicollinearity from the data uses only information from the data corresponding to the predictors and we do not consider the relation to the response. Developing a transformation processing method that incorporates information from both types of data could further improve the variable selection performance. Although numerical examples confirmed that the performance of TPPIS is better than that of existing methods, no mathematical proof is provided. In the process of devising a proof, we may be able to identify the characteristics of the data for which TPPIS is most effective.

Acknowledgement

The authors are grateful to the anonymous reviewer for valuable comments and suggestions that improve the quality of this paper. This work was supported by JSPS KAKENHI Grant Numbers 19K11858 and 23K11005.

References

Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics; 2013 29 April- 1 May*, pages 126–134. Scottsdale, Arizona. PMLR.

Chowdhury, M. Z. I. and Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8(1).

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer.

[1] https://fred.stlouisfed.org/series/SP500

[2] https://www.kaggle.com/hanseopark/sp-500-stocks-value-with-financial-statement

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

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. (2018). Sure independence screening. *Wiley StatsRef: Statistics Reference Online*.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, 38:3567–3604.

Fang, T., Lee, T.-H., and Su, Z. (2020). Predicting the long-term stock market volatility: A GARCH-MIDAS model with variable selection. *Journal of Empirical Finance*, 58:36–49.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction (2nd ed.)*. Springer.

Helwig, N., Pignanelli, E., and Schütze, A. (2015). Condition monitoring of a complex hydraulic system using multivariate statistics. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings; 2015 11-14 May*, pages 210–215. Pisa, Italy. IEEE.

Jia, J. and Rohe, K. (2012). Preconditioning to comply with the irrepresentable condition. *arXiv preprint arXiv:1208.5584*.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *Annals of Statistics*, 40:1846–1877.

Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Tian, S., Yu, Y., and Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100.

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

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, 99(1):15–28.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):589–611.

Yun, Y.-H., Li, H.-D., Deng, B.-C., and Cao, D.-S. (2019). An overview of variable selection methods in multivariate analysis of near-infrared spectra. *Trends in Analytical Chemistry*, 113:102–115.

Zhang, J., Liu, Y., and Wu, Y. (2017). Correlation rank screening for ultrahigh-dimensional survival data. *Computational Statistics & Data Analysis*, 108:121–132.

Zhao, N., Xu, Q., Tang, M.-L., Jiang, B., Chen, Z., and Wang, H. (2020). High-dimensional variable screening under multicollinearity. *Stat*, 9(1):e272.

Received: May 11, 2024
Revised: September 10, 2024
Accept: October 21, 2024

Table 1: Simulation results for Example 1

n	p	φ	Method	best d	best α	BIC	F2-score	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$
100	1000	0.5	SIS	-	-	7.982	0.285	38	27	32	0
			FPSIS	1	-	5.791	0.914	91	90	95	88
			FPSIS$_d$	20	-	5.736	0.940	96	95	94	93
			PPIS	1	-	5.674	0.948	96	95	95	93
			TPPIS$_\alpha$	1	0.8	5.590	0.973	98	97	99	97
			TPPIS$_{d,\alpha}$	1	0.8	5.590	0.973	98	97	99	97
		0.7	SIS	-	-	7.583	0.283	38	35	24	0
			FPSIS	1	-	5.691	0.940	95	93	96	92
			FPSIS$_d$	20	-	5.643	0.964	97	96	98	96
			PPIS	1	-	5.632	0.962	98	97	95	95
			TPPIS$_\alpha$	1	1.0	5.632	0.962	98	97	95	95
			TPPIS$_{d,\alpha}$	20	1.0	5.623	0.971	100	97	97	97
		0.9	SIS	-	-	6.664	0.250	26	30	30	0
			FPSIS	1	-	5.567	0.969	97	99	96	96
			FPSIS$_d$	1	-	5.567	0.969	97	99	96	96
			PPIS	1	-	5.550	0.984	98	100	98	98
			TPPIS$_\alpha$	1	1.0	5.550	0.984	98	100	98	98
			TPPIS$_{d,\alpha}$	1	1.0	5.550	0.984	98	100	98	98
300	1000	0.5	SIS	-	-	9.009	0.462	61	66	67	0
			FPSIS	1	-	6.200	0.981	100	100	100	97
			FPSIS$_d$	1	-	6.200	0.981	100	100	100	97
			PPIS	1	-	6.137	0.993	100	100	100	98
			TPPIS$_\alpha$	1	1.0	6.137	0.993	100	100	100	98
			TPPIS$_{d,\alpha}$	1	1.0	6.137	0.993	100	100	100	98
		0.7	SIS	-	-	8.590	0.476	72	66	67	0
			FPSIS	1	-	6.252	0.971	100	100	100	95
			FPSIS$_d$	60	-	6.247	0.968	98	99	99	94
			PPIS	1	-	6.323	0.962	100	99	99	92
			TPPIS$_\alpha$	1	0.8	6.206	0.977	99	99	100	96
			TPPIS$_{d,\alpha}$	60	0.6	6.113	0.994	100	100	100	99
		0.9	SIS	-	-	7.722	0.365	48	46	53	0
			FPSIS	1	-	6.168	0.978	100	100	100	95
			FPSIS$_d$	60	-	6.167	0.978	99	100	99	95
			PPIS	1	-	6.123	0.987	99	99	100	97
			TPPIS$_\alpha$	1	0.8	6.086	0.996	100	100	100	99
			TPPIS$_{d,\alpha}$	1	0.8	6.086	0.996	100	100	100	99

Table 2: Simulation results for Example 2

n	p	φ	Method	best d	best α	BIC	F2-score	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$
100	1000	0.5	SIS	-	-	8.229	0.264	2	1	9	0	100
			FPSIS	1	-	6.420	0.915	95	93	93	94	97
			FPSIS$_d$	1	-	6.420	0.915	95	93	93	94	97
			PPIS	1	-	6.176	0.933	93	97	96	92	96
			TPPIS$_\alpha$	1	1.0	6.176	0.933	93	97	96	92	96
			TPPIS$_{d,\alpha}$	1	1.0	6.176	0.933	93	97	96	92	96
		0.7	SIS	-	-	7.736	0.238	0	0	0	0	100
			FPSIS	1	-	6.814	0.845	78	78	88	91	99
			FPSIS$_d$	1	-	6.814	0.845	78	78	88	91	99
			PPIS	1	-	6.344	0.920	92	90	93	93	99
			TPPIS$_\alpha$	1	1.0	6.344	0.920	92	90	93	93	99
			TPPIS$_{d,\alpha}$	1	1.0	6.344	0.920	92	90	93	93	99
		0.9	SIS	-	-	6.714	0.238	0	0	0	0	100
			FPSIS	1	-	7.908	0.343	21	17	19	83	29
			FPSIS$_d$	20	-	6.273	0.893	88	85	89	96	98
			PPIS	1	-	6.214	0.904	90	85	90	92	100
			TPPIS$_\alpha$	1	1.0	6.214	0.904	90	85	90	92	100
			TPPIS$_{d,\alpha}$	1	1.0	6.214	0.904	90	85	90	92	100
300	1000	0.5	SIS	-	-	9.182	0.584	65	67	69	0	100
			FPSIS	1	-	6.378	0.983	100	100	100	94	100
			FPSIS$_d$	1	-	6.378	0.983	100	100	100	94	100
			PPIS	1	-	6.367	0.983	100	100	100	95	100
			TPPIS$_\alpha$	1	0.8	6.362	0.985	100	100	100	95	100
			TPPIS$_{d,\alpha}$	1	0.8	6.362	0.985	100	100	100	95	100
		0.7	SIS	-	-	8.781	0.491	47	58	43	0	100
			FPSIS	1	-	6.313	0.988	100	100	100	96	100
			FPSIS$_d$	1	-	6.313	0.988	100	100	100	96	100
			PPIS	1	-	6.327	0.985	100	100	100	96	100
			TPPIS$_\alpha$	1	0.6	6.271	0.990	100	99	99	98	100
			TPPIS$_{d,\alpha}$	1	0.6	6.271	0.990	100	99	99	98	100
		0.9	SIS	-	-	7.829	0.269	6	6	7	0	100
			FPSIS	1	-	6.379	0.960	99	100	100	97	100
			FPSIS$_d$	60	-	6.301	0.978	99	99	99	95	100
			PPIS	1	-	6.279	0.981	99	99	99	97	100
			TPPIS$_\alpha$	1	0.8	6.240	0.990	99	99	99	98	100
			TPPIS$_{d,\alpha}$	1	0.8	6.240	0.990	99	99	99	98	100

Table 3: Simulation results for Example 3

n	p	φ	Method	best d	best α	BIC	F2-score	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$
100	1000	0.5	SIS	-	-	8.223	0.178	2	0	0	0	73	26
			FPSIS	1	-	6.735	0.824	87	85	89	85	88	75
			FPSIS$_d$	20	-	6.649	0.848	86	91	88	88	88	70
			PPIS	1	-	6.331	0.904	94	95	94	93	94	79
			TPPIS$_\alpha$	1	1.0	6.331	0.904	94	95	94	93	94	79
			TPPIS$_{d,\alpha}$	1	1.0	6.331	0.904	94	95	94	93	94	79
		0.7	SIS	-	-	7.733	0.164	0	0	0	0	69	31
			FPSIS	1	-	6.928	0.786	83	81	74	88	87	66
			FPSIS$_d$	20	-	6.511	0.868	90	92	90	93	88	65
			PPIS	1	-	6.275	0.899	96	92	93	92	93	80
			TPPIS$_\alpha$	1	1.0	6.275	0.899	96	92	93	92	93	80
			TPPIS$_{d,\alpha}$	1	1.0	6.275	0.899	96	92	93	92	93	80
		0.9	SIS	-	-	6.683	0.202	0	0	0	0	85	15
			FPSIS	1	-	7.831	0.350	20	14	16	89	29	6
			FPSIS$_d$	20	-	6.517	0.808	80	79	83	93	79	34
			PPIS	1	-	6.191	0.928	92	96	96	97	98	70
			TPPIS$_\alpha$	1	1.0	6.191	0.928	92	96	96	97	98	70
			TPPIS$_{d,\alpha}$	1	1.0	6.191	0.928	92	96	96	97	98	70
300	1000	0.5	SIS	-	-	9.275	0.447	50	48	54	0	93	65
			FPSIS	1	-	6.592	0.924	97	97	100	92	98	89
			FPSIS$_d$	60	-	6.499	0.947	99	97	100	93	99	76
			PPIS	1	-	6.359	0.955	99	99	98	98	99	79
			TPPIS$_\alpha$	1	1.0	6.359	0.955	99	99	98	98	99	79
			TPPIS$_{d,\alpha}$	1	1.0	6.359	0.955	99	99	98	98	99	79
		0.7	SIS	-	-	8.834	0.316	20	21	21	0	94	28
			FPSIS	1	-	6.491	0.942	99	99	99	94	99	90
			FPSIS$_d$	60	-	6.462	0.948	98	97	98	94	100	55
			PPIS	1	-	6.502	0.943	100	98	99	94	100	72
			TPPIS$_\alpha$	1	0.4	6.393	0.952	99	100	99	98	99	88
			TPPIS$_{d,\alpha}$	1	0.4	6.393	0.952	99	100	99	98	99	88
		0.9	SIS	-	-	7.830	0.237	1	0	0	0	99	2
			FPSIS	1	-	6.463	0.936	97	99	97	98	100	80
			FPSIS$_d$	60	-	6.354	0.962	97	97	97	97	96	7
			PPIS	1	-	6.315	0.969	99	99	99	98	100	52
			TPPIS$_\alpha$	1	1.0	6.315	0.969	99	99	99	98	100	52
			TPPIS$_{d,\alpha}$	60	1.0	6.260	0.989	100	99	99	99	99	4

Table 4: Simulation results for Example 4

n	p	d	m	Method	best d	best α	BIC	F2-score	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$
100	1000	3	20	SIS	-	-	10.537	0.377	100	25	10	0
				FPSIS	3	-	10.961	0.285	34	67	11	0
				FPSIS$_d$	20	-	10.865	0.338	52	65	13	5
				PPIS	3	-	10.385	0.578	88	93	45	4
				TPPIS$_\alpha$	3	1.0	10.385	0.578	88	93	45	4
				TPPIS$_{d,\alpha}$	3	1.0	10.385	0.578	88	93	45	4
			40	SIS	-	-	10.858	0.488	100	44	37	0
				FPSIS	3	-	11.613	0.213	18	60	5	0
				FPSIS$_d$	20	-	11.219	0.468	78	83	21	5
				PPIS	3	-	10.618	0.676	99	98	62	10
				TPPIS$_\alpha$	3	1.0	10.618	0.676	99	98	62	10
				TPPIS$_{d,\alpha}$	3	1.0	10.618	0.676	99	98	62	10
			60	SIS	-	-	11.129	0.563	100	58	56	0
				FPSIS	3	-	11.790	0.337	40	72	22	0
				FPSIS$_d$	20	-	11.506	0.470	80	76	21	6
				PPIS	3	-	10.852	0.738	100	98	76	19
				TPPIS$_\alpha$	3	1.0	10.852	0.738	100	98	76	19
				TPPIS$_{d,\alpha}$	3	1.0	10.852	0.738	100	98	76	19
			80	SIS	-	-	11.443	0.611	100	57	78	0
				FPSIS	3	-	11.835	0.501	74	73	55	0
				FPSIS$_d$	3	-	11.835	0.501	74	73	55	0
				PPIS	3	-	11.180	0.753	99	99	87	9
				TPPIS$_\alpha$	3	1.0	11.180	0.753	99	99	87	9
				TPPIS$_{d,\alpha}$	3	1.0	11.180	0.753	99	99	87	9
300	1000	3	60	SIS	-	-	11.571	0.669	100	89	94	0
				FPSIS	3	-	11.726	0.604	99	100	78	0
				FPSIS$_d$	60	-	10.464	0.929	100	100	100	89
				PPIS	3	-	10.279	0.981	100	100	100	100
				TPPIS$_\alpha$	3	1.0	10.279	0.981	100	100	100	100
				TPPIS$_{d,\alpha}$	3	1.0	10.279	0.981	100	100	100	100
			120	SIS	-	-	12.268	0.758	100	100	100	0
				FPSIS	3	-	12.265	0.772	100	100	100	0
				FPSIS$_d$	60	-	11.867	0.890	100	100	99	69
				PPIS	3	-	11.692	0.945	100	100	100	83
				TPPIS$_\alpha$	3	1.0	11.692	0.945	100	100	100	83
				TPPIS$_{d,\alpha}$	3	1.0	11.692	0.945	100	100	100	83
			180	SIS	-	-	12.951	0.772	100	100	100	0
				FPSIS	3	-	12.926	0.784	100	100	100	0
				FPSIS$_d$	60	-	12.923	0.795	100	99	90	35
				PPIS	3	-	12.708	0.884	100	100	100	59
				TPPIS$_\alpha$	3	1.0	12.708	0.884	100	100	100	59
				TPPIS$_{d,\alpha}$	3	1.0	12.708	0.884	100	100	100	59
			240	SIS	-	-	13.481	0.780	100	100	100	0
				FPSIS	3	-	13.469	0.785	100	100	100	0
				FPSIS$_d$	3	-	13.469	0.785	100	100	100	0
				PPIS	3	-	13.477	0.796	100	100	88	28
				TPPIS$_\alpha$	3	0.8	13.435	0.811	100	100	93	27
				TPPIS$_{d,\alpha}$	3	0.8	13.435	0.811	100	100	93	27

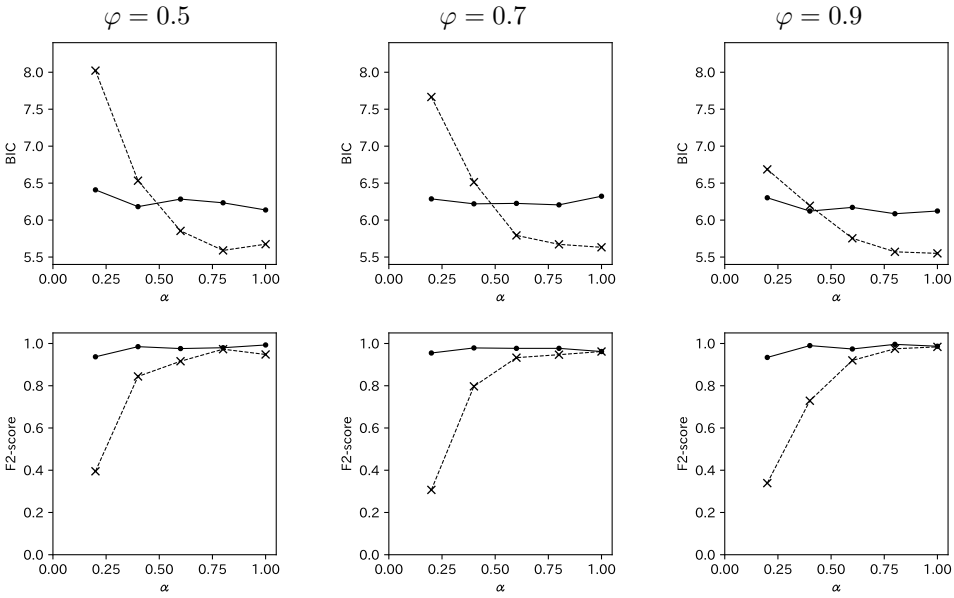

Figure 1: Values of BIC and F2-score for different α in TPPIS of Example 1. The top row shows BIC results and the bottom row shows F2-score results. The values for $n = 300$ are represented by $\bullet$, and the values for $n = 100$ are represented by $\times$. p is 1000 in all cases.

Table 5: Results for condition monitoring of hydraulic systems

Method	best d	best α	BIC	Number of selected variables
SIS	-	-	11.740	12
FPSIS	2	-	12.452	8
FPSIS_d	2	-	12.452	8
PPIS	2	-	12.491	41
TPPIS_α	2	0.4	11.711	17
$\text{TPPIS}_{d,\alpha}$	2	0.4	11.711	17

Table 6: Results for S&P500

Method	best d	best α	BIC	Number of selected variables
SIS	-	-	13.646	2
FPSIS	2	-	13.513	5
FPSIS_d	101	-	12.934	17
PPIS	2	-	13.346	11
TPPIS_α	2	0.8	13.119	15
$\text{TPPIS}_{d,\alpha}$	50	0.4	12.629	7