Sparse Hilbert–Schmidt Independence Criterion Regression

Benjamin Poignard
Osaka University/RIKEN AIP
bpoignard@econ.osaka-u.ac.jp

Makoto Yamada
RIKEN AIP/Kyoto University/JST PRESTO
makoto.yamada@riken.jp

Abstract

Feature selection is a fundamental problem for machine learning and statistics, and it has been widely studied over the past decades. However, the majority of feature selection algorithms are based on linear models, and the nonlinear feature selection problem has not been well studied compared to linear models, in particular for the high-dimensional case. In this paper, we propose the sparse Hilbert–Schmidt Independence Criterion (SpHSIC) regression, which is a versatile nonlinear feature selection algorithm based on the HSIC and is a continuous optimization variant of the well-known minimum redundancy maximum relevance (mRMR) feature selection algorithm. More specifically, the SpHSIC consists of two parts: the convex HSIC loss function on the one hand and the regularization term on the other hand, where we consider the Lasso, Bridge, MCP, and SCAD penalties. We prove that the sparsity based HSIC regression estimator satisfies the oracle property; that is, the sparsity-based estimator recovers the true underlying sparse model and is asymptotically normally distributed. On the basis of synthetic and real-world experiments, we illustrate this theoretical property and highlight the fact that the proposed algorithm performs well in the high-dimensional setting.

1 Introduction

Feature selection/variable selection, which consists of selecting a subset of features in high-dimensional data, is a widely studied problem in the machine learning and statistics communities, and has many real-world applications, including, *e.g.*, biomarker discovery from expression data (Peng *et al.*, 2005; Yamada *et al.*, 2018), microarray gene expression data classification (Abusamra, 2013), and text mining (Forman, 2003).

One standard feature selection approach is based on sparse modeling, where the least absolute shrinkage and selection operator (Lasso) is probably the most commonly used *linear* feature selection algorithm (Tibshirani, 1996). Lasso has been applied to a broad range of applications, and its theoretical properties such as the consistency and the conditions for support recovery – in the adaptive case – have been widely studied (Hastie *et al.*, 2015). Most of the applications of Lasso concern the standard linear regression model. It often provides low performance when the underlying data cannot be represented through a linear model. For example, in gene expression data, the data generation process is, in general, unknown; using a linear model may not be the best choice.

To handle complex data, *nonlinear* modeling offers a more suitable alternative. One of the widely used nonlinear feature selection algorithms is the sparse additive model (SpAM) (Ravikumar *et al.*, 2009) and its variants, where SpAM is a sparse variant of the well-known additive model. SpAM outperforms linear models including Lasso in various setups (Ravikumar *et al.*, 2009) and its theoretical properties are also established. However, SpAM assumes the additiveness of functions with univariate inputs, and the prediction performance can be significantly lowered if the additive assumption is violated (*e.g.,* the multiplicative model).

Another standard nonlinear feature selection approach is based on the *screening* method, in which the important features are selected by ranking the association measure between each feature and its corresponding output (Fan and Lv, 2008; Balasubramanian *et al.*, 2013). After the screening step, a non-parametric model is fitted to the screened features. Because this screening-based approach simply selects features if a relationship exists between a feature and the output,

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

strong assumptions, such as additiveness in SpAM, are not required. Moreover, the sure screening property can be proved, which ensures that the screening method can select true features with high probability (Fan and Lv, 2008; Balasubramanian *et al.*, 2013).

Typically, the mutual information (MI) (Cover and Thomas, 2006), distance correlation (Székely *et al.*, 2009; Li *et al.*, 2012), and the Hilbert–Schmidt independence criterion (HSIC) (Gretton *et al.*, 2005; Balasubramanian *et al.*, 2013) are used as an association measure of the screening method. One of the limitations of screening methods is that they tend to select redundant features when a significant number of similar features are related to the output, which thus may lower the prediction performance.

The minimum redundancy maximum relevance (mRMR) feature selection (Peng *et al.*, 2005) can be a good candidate for a screening method to deal with the redundancy problem. mRMR has been well studied in the data mining community, and some studies showed that it outperforms the linear models: see, *e.g.*, (Haws *et al.*, 2015). However, to the best of our knowledge, the theoretical properties of mRMR methods and their variants have not been established.

In this paper, we propose the sparse Hilbert–Schmidt Independence Criterion regression (SpHSIC) together with a large sample analysis of the mRMR approach. More specifically, we first consider the continuous optimization variant of the mRMR algorithm, in which the loss function can be represented by the difference between the centered input Gram matrices and the centered output Gram matrices, which then provides a regression model representation based on V-statistics linking the dependent variable with the set of features through a linear combination. Under the sparsity assumption, we propose a penalization framework to recover the true sparse support, that is, the key features, where the set of penalties is given by Lasso (Tibshirani, 1996), Bridge (Frank and Friedman, 1993), SCAD (Fan and Li, 2001), and MCP (Zhang *et al.*, 2010), which all are non-convex except Lasso.

The first key contribution of this paper is to propose the SpHSIC regression framework, which is a continuous optimization variant of the mRMR algorithm. The second contribution is to present the asymptotic theory of the sparsity-based HSIC estimator. We prove the oracle property in the sense of (Fan and Li, 2001) for non-convex penalties; that is, the sparsity-based estimator recovers the true underlying sparse model and is asymptotically normally distributed. The third contribution is to conduct a large sample analysis for V-statistics-based loss functions. Using the asymptotic equivalence between V-statistics and U-statistics, we rewrite the V-statistic-based loss as a U-statistic-based loss with a symmetric kernel of degree four. A key assumption to derive asymptotic results is the non-independence between the dependent variable and the features, which allows for working with non-degenerate U-statistics. Our study shares a similar spirit with (Rejchel, 2017), who derived some large sample properties for degree two U-statistics based and convex loss function with an adaptive Lasso type penalty. But our work differs from the latter in two main respects: the loss function involves a degree four kernel function requiring a careful treatment of its U-statistics representation and its degeneracy; we consider a general framework encompassing a broad range of potentially non-convex penalty functions.

2 Preliminaries

We first briefly review the framework of the HSIC that will be used throughout this paper. More details can be found in (Gretton *et al.*, 2005) or (Song *et al.*, 2012).

2.1 Problem formulation

Let $\mathcal{X}$ be a metric space and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) induced by the inner product $\langle .,. \rangle$ if there exists a function $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$\text{(i) } \forall \boldsymbol{x} \in \mathcal{X},\ \phi(\boldsymbol{x}, .) \in \mathcal{H},$$
$$\text{(ii) } \forall f \in \mathcal{H}, \forall \boldsymbol{x} \in \mathcal{X}, \langle f, \phi(\boldsymbol{x}, .)\rangle = f(\boldsymbol{x}).$$

For any probability measure $\mathbb{P}$ defined on $\mathcal{X}$, the mean $\mu(\mathbb{P}) \in \mathcal{H}$ is defined as $\mathbb{E}[f(X)] = \langle f, \mu(\mathbb{P})\rangle$ for any $f \in \mathcal{H}$ with X sampled from $\mathcal{X}$.

For the formal setting, we consider two random variables $Y \sim \mathbb{P}_Y$ and $X \sim \mathbb{P}_X$ that take values on $(\mathcal{Y}, \mathcal{B}_y)$ and $(\mathcal{X}, \mathcal{B}_x)$, respectively, where $\mathcal{Y}, \mathcal{X}$ are two separable metrics, and $\mathcal{B}_y, \mathcal{B}_x$ are Borel σ-algebras. Then, $(\mathcal{Y} \times \mathcal{X}, \mathcal{B}_y \times \mathcal{B}_x)$ is measurable, and the joint distribution is defined as $\mathbb{P}_{Y\,X}$, which assigns values to the product space $(\mathcal{Y} \times \mathcal{X}, \mathcal{B}_y \times \mathcal{B}_x)$. We then define the symmetric kernels $\phi(.,.)$ and $\psi(.,.)$ on the spaces $\mathcal{Y}$ and $\mathcal{X}$, respectively, and assume $\mathbb{E}_Y[\phi(Y,Y)] < \infty$ and $\mathbb{E}_X[\psi(X,X)] < \infty$.

The objective of this paper is to provide a procedure for selecting a subset of the features X that are important for its output Y. We suppose that we observe n samples $\{(Y_1, X_1), \cdots, (Y_n, X_n)\}$ from $(\mathcal{Y} \times \mathcal{X}, \mathcal{B}_y \times \mathcal{B}_x)$.

2.2 Association-based feature selection

The simplest association-based feature selection algorithm would be based on maximum relevance (MR) feature selection (Peng *et al.*, 2005):

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{\text{argmax}} \quad \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} D(X^{(k)}, Y),$$

where $D(X^{(k)}, Y) \geq 0$ is the association score between the k-th feature and its output. In the original paper, the MI is used as the association score of the MR method (Peng *et al.*, 2005). Moreover, this MR method is related to the sure independence screening method (Fan and Lv, 2008), which guarantees selection of true features with high probability.

However, the MR method tends to select redundant features (*i.e.*, the selected features can be highly correlated), because it does not use the feature-feature relationship to select features. To deal with this problem, the mRMR feature selection algorithm was developed by (Peng *et al.*, 2005), and the objective function can be reformulated as

$$\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} D(X^{(k)}, Y) - \frac{1}{|\mathcal{S}|^2}\sum_{k\in\mathcal{S}}\sum_{k'\in\mathcal{S}} D(X^{(k)}, X^{(k')}), \quad (1)$$

where $\mathcal{S}$ is the set of selected feature indices. The second term in Eq. (1) is the penalized term that selects independent features. More specifically, since $D(X^{(k)}, X^{(k')})$ takes nonnegative values if $X^{(k)}$ and $X^{(k')}$ are non-independent, the second term takes large negative values if the selected features are not mutually independent. Thus, selecting features by maximizing Eq. (1), we can select features that are dependent on the output, and the selected features are mutually independent. Although some studies provide empirical results which highlight that the mRMR algorithm performs well in practice, the theoretical properties of the mRMR method have not been studied.

2.3 Hilbert–Schmidt independence criterion

Here, we review the HSIC. More details can be found in (Gretton *et al.*, 2005). The HSIC of $\mathbb{P}_{YX}$ is

$$\begin{aligned}\text{HSIC}(Y, X) &= \mathbb{E}_{YY'XX'}[\phi(Y, Y')\psi(X, X')]\\ &+ \mathbb{E}_{YY'}[\phi(Y, Y')]\mathbb{E}_{XX'}[\psi(X, X')]\\ &- 2\mathbb{E}_{YX}[\mathbb{E}_{Y'}[(\phi(Y, Y'))]\mathbb{E}_{X'}[\psi(X, X')]],\end{aligned}$$

where (Y', X') is an i.i.d. copy of (Y, X), and $\mathbb{E}_{XX'}[.]$ (resp. $\mathbb{E}_X[.]$) is the expectation defined over X, X' (resp. X). Following the V-statistic-based HSIC estimator of Gretton *et al.* (2005), we define

$$\widehat{\text{HSIC}}(Y, X) = \frac{1}{n^2}\text{trace}(\boldsymbol{L}\,\boldsymbol{H}_n\,\boldsymbol{K}\,\boldsymbol{H}_n), \quad (2)$$

where $L_{ij} = \phi(Y_i, Y_j)$ and $K_{ij} = \psi(X_i, X_j)$ are kernel functions, assumed symmetric; $\boldsymbol{L} = (L_{ij}) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{K} = (K_{ij}) \in \mathbb{R}^{n\times n}$ are Gram matrices; $\boldsymbol{H}_n = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^\top$ is a centering matrix; $\boldsymbol{I}_n$ is the $n \times n$ dimensional identity matrix; $\boldsymbol{1}_n$ is the n-dimensional vector whose elements are all 1; and $^\top$ denotes the transpose.

Throughout this paper, we define $\boldsymbol{Z}_i = (Y_i, X_i)$, where the random variable Y is of size p and X of size q, and a random vector containing d features is denoted as $X_i^{(1)}, \cdots, X_i^{(d)}$ for any observation i.

2.4 HSIC Lasso

We now briefly review the HSIC Lasso originally proposed by (Yamada *et al.*, 2014). The original mRMR algorithm consists of a discrete optimization problem, and the optimization is in general difficult. To mitigate the problem, continuous optimization tends to be used. We first rewrite the mRMR algorithm as

$$\begin{aligned}\underset{\boldsymbol{\beta}\in\{0,1\}^d}{\text{argmax}} \quad &\frac{1}{\boldsymbol{\beta}^\top\boldsymbol{1}_d}\sum_{k=1}^{d}\beta_k\widehat{\text{HSIC}}(X^{(k)}, Y)\\ &- \frac{1}{(\boldsymbol{\beta}^\top\boldsymbol{1}_d)^2}\sum_{k=1}^{d}\sum_{k'=1}^{d}\beta_k\beta_{k'}\widehat{\text{HSIC}}(X^{(k)}, X^{(k')}).\end{aligned}$$

where $\boldsymbol{1}_d \in \mathbb{R}^d$ is the vector whose elements are all one. We then relax $\boldsymbol{\beta}$ as $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^\top \in \mathbb{R}^d$. Moreover, because $\boldsymbol{\beta}$ is a sparse vector, we can write the relaxed version of the optimization problem as

$$\begin{aligned}\underset{\boldsymbol{\theta}\in\mathbb{R}_+^d}{\text{argmax}} \quad &\sum_{k=1}^{d}\theta_k\widehat{\text{HSIC}}(X^{(k)}, Y)\\ &- \frac{1}{2}\sum_{k=1}^{d}\sum_{k'=1}^{d}\theta_k\theta_{k'}\widehat{\text{HSIC}}(X^{(k)}, X^{(k')}) - \lambda\|\boldsymbol{\theta}\|_1,\end{aligned}$$

where $\|\boldsymbol{\theta}\|_1$ is the ℓ_1 norm, and $\lambda \geq 0$ is the regularization parameter. Note that the nonnegative constraint is added, because the original $\boldsymbol{\beta}$ is nonnegative.

Here, we simply drop the sum to one constraint, because it makes the problem easy to solve. This optimization problem is convex. However, it needs to compute a $d \times d$ dimensional HSIC matrix, which is computationally expensive. Moreover, this formulation is not appropriate for a high-dimensional setup, because it requires $O(d^2)$ memory space. To deal with the problem, we express the optimization problem as

$$\underset{\boldsymbol{\theta}\in\mathbb{R}_+^d}{\text{argmin}} \quad \|\bar{\boldsymbol{L}} - \sum_{k=1}^{d}\theta_k\bar{\boldsymbol{K}}^{(k)}\|_F^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

where $\bar{\boldsymbol{L}} = \boldsymbol{H}_n\boldsymbol{L}\boldsymbol{H}_n$ is the centered Gram matrix of Y, with $L_{ij} = \phi(Y_i, Y_j)$; $\bar{\boldsymbol{K}}^{(k)} = \boldsymbol{H}_n\boldsymbol{K}^{(k)}\boldsymbol{H}_n$ is the centered Gram matrix of the k-th input $X^{(k)}$; and $\boldsymbol{K}^{(k)} = (K_{ij}^{(k)}) \in \mathbb{R}^{n\times n}$ with $K_{ij}^{(k)} = \psi(X_i^{(k)}, X_j^{(k)})$ is the input Gram matrix. This formulation makes the required memory space for the Gram matrices $O(dn^2)$, which is more appropriate for a high-dimensional setting (*i.e.*, $n \ll d$).

3 SpHSIC Regression

In this section, we present the SpHSIC regression. Following the above development of the HSIC Lasso formulation, we consider the regression model:

$$\forall i, j, \left(\bar{\boldsymbol{L}}\right)_{ij} = \sum_{k=1}^{d} \theta_k \left(\bar{\boldsymbol{K}}^k\right)_{ij} + U_{ij}, \tag{3}$$

where U_{ij} is a centered error term with $\mathbb{E}[U_{ij}^2] = \sigma^2$.

The sparse HSIC regression (SpHSIC) (a.k.a., V-statistic based M-estimation criterion) is defined as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left\{ \mathbb{G}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) + \sum_{k=1}^{d} \varphi\left(\frac{\lambda_n}{n}, |\theta_k|\right) \right\},$$

with

$$\mathbb{G}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) := \frac{1}{n^2} \sum_{i,j=1}^{n} \left(\left(\bar{\boldsymbol{L}}\right)_{ij} - \sum_{k=1}^{d} \theta_k \left(\bar{\boldsymbol{K}}^{(k)}\right)_{ij} \right)^2,$$

where

$$\left(\bar{\boldsymbol{L}}\right)_{ij} = L_{ij} - \frac{1}{n} \sum_{j'=1}^{n} L_{ij'} - \frac{1}{n} \sum_{i'=1}^{n} L_{i'j} + \frac{1}{n^2} \sum_{i'=1}^{n} \sum_{j'=1}^{n} L_{i'j'},$$

$\forall\ i, j \leq n$, and $\bar{\boldsymbol{K}}^{(k)}$ is defined in the same way based on $K_{ij}^{(k)} = \psi(X_i^{(k)}, X_j^{(k)})$. SpHSIC is based on the transformed data L_{ij} and K_{ij}^k, which are centered by the empirical mean over all the components of L and the column-wise and row-wise means.

The penalization is performed through the term $\sum\limits_{k=1}^{d} \varphi\left(\frac{\lambda_n}{n}, |\theta_k|\right)$, which is a coordinate-separable penalty. We consider the following set of penalties:

$$\begin{array}{lll}
\textbf{Lasso} & : & \lambda|\theta|, \\
\textbf{Bridge} & : & \lambda|\theta|^q, q \in (0,1), \\
\textbf{MCP} & : & \operatorname{sgn}(\theta)\lambda \int_0^{|\theta|} (1 - z/(\lambda b_{\text{mcp}}))_+ \mathrm{d}z, \\
\textbf{SCAD} & : & \begin{cases} \lambda|\theta|, & \text{for } |\theta| \leq \lambda, \\ -\frac{(\theta^2 - 2b_{\text{scad}}\lambda|\theta| + \lambda^2)}{(2(b_{\text{scad}} - 1))}, & \text{for } \lambda \leq |\theta| \leq b_{\text{scad}}\lambda, \\ (b_{\text{scad}} + 1)\lambda^2/2, & \text{for } |\theta| > b_{\text{scad}}\lambda. \end{cases}
\end{array}$$

and $\lambda \geq 0$ is the regularization parameter.

Note that the HSIC Lasso (Yamada *et al.*, 2014) corresponds to the SpHSIC with the ℓ_1 regularizer. In this paper, we consider Bridge, MCP, and SCAD in addition to the ℓ_1 regularizer. For the ℓ_1 regularizer, the objective function is convex. In contrast, the penalized framework SpHSIC that we proposed using other penalty functions is non-convex.

4 Theoretical Analysis

In this section, we provide a theoretical analysis of the SpHSIC.

4.1 Parameter-dependent U-statistics

The non-penalized loss function $\mathbb{G}_n(.)$ is a V-statistic-based loss function. Because we develop our analysis asymptotically, we will use the U-statistic framework rather than the V-statistic one. We thus propose to rewrite the non-penalized V-statistic-based loss $\mathbb{G}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})$ as a U-statistic $\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})$, which is a summation encompassing single indices and pairs, triplets and quadruplets of indices, and develop our large sample analysis using this loss. Indeed, there is a $\sqrt{n}$-asymptotic equivalence between the two so that the large sample distribution of the U-statistic is the same as that of the V-statistic: see subsection 5.7.3 of (Serfling, 1980) and subsection 4.2 of (Lee, 1990) for further details. Then, following the same reasoning as (Song *et al.*, 2012), where the HSIC statistics can be expressed as a U-statistic with a symmetric kernel of degree 4 (see their Theorem 3), we may express the least squares criterion as a U-statistic with a symmetric kernel of degree 4.

Proposition 4.1. *The non-penalized loss function $\mathbb{G}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})$ can be rewritten in terms of a U-statistic $\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})$ as*

$$\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) = (n)_4^{-1} \sum_{(i,j,q,r) \in \boldsymbol{I}_4^n} \ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}),$$

where $(n)_c = \frac{n!}{(n-c)!}$ and

$$\begin{aligned}
&\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}) \\
&= \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} \{ \big(L_{st} - \sum_{k=1}^{d} \theta_k K_{st}^k\big)\big(L_{st} - \sum_{l=1}^{d} \theta_l K_{st}^k\big) \\
&\quad + \big(L_{st} - \sum_{k=1}^{d} \theta_k K_{st}^k\big)\big(L_{uv} - \sum_{l=1}^{d} \theta_l K_{uv}^k\big) \\
&\quad + \sum_{k=1}^{d} \theta_k K_{uv}^k L_{st} - \sum_{k=1}^{d} \theta_k K_{st}^k L_{uv} \\
&\quad - 2\big[\big(L_{st} - \sum_{k=1}^{d} \theta_k K_{st}^k\big)\big(L_{su} - \sum_{l=1}^{d} \theta_l K_{su}^l\big) \\
&\quad + \sum_{k=1}^{d} \theta_k K_{su}^k L_{st} - \sum_{k=1}^{d} \theta_k K_{st}^k L_{su}\big] \},
\end{aligned} \tag{4}$$

is the symetrized kernel. The sum is taken over all ordered quadruples (s, t, u, v) selected without replacement from (i, j, q, r), and $\boldsymbol{I}_4^n$ denotes the set of all 4tuples drawn without replacement from $\{1, \cdots, n\}$.

Remark. The proof of this proposition relies on the use of the U-statistic-based expression of the estimator of the HSIC(Y, X) derived by (Song *et al.*, 2012). It can be found in the supplementary file.

This expression will enable us to use limit theorems for parameter-dependent U-statistics. Indeed, the score vector $\nabla_{\boldsymbol{\theta}_0}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)$ ($\boldsymbol{\theta}_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,d})^\top$) is formed by U-statistics, which are assumed to be non-degenerate. A condition for the elements of the score-based U-statistic not being first order degenerate is that

$$\mathrm{Var}(\tilde{\ell}(\boldsymbol{Z}_i, \boldsymbol{\theta}_0)) \neq 0,$$

with $\tilde{\ell}(\boldsymbol{Z}_i; \boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{Z}_j\boldsymbol{Z}_q\boldsymbol{Z}_r}[\nabla_{\theta_k}\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)]$, where the latter expectation corresponds to an integration with respect to $\boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r$. Assumption 2 is required to avoid an order one degenerate score-based U-statistic, which is crucial for the existence of the positive-definite variance covariance matrix

$$\mathbb{M}(\boldsymbol{\theta}_0) = \mathbb{E}[\nabla_\theta\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)\nabla_{\theta^\top}\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)],$$

which will serve as the asymptotic variance covariance matrix when applying the multivariate central limit theorem for U-statistics. This assumption is key when deriving the limiting distribution because the distribution will be Gaussian under this assumption. On the contrary, if the variables are independent and thus each component of the U-statistic-based score function is first order degenerate, each of these elements' distribution would be an infinite sum of the $\chi^2(1)$ distribution. This is the motivation for the following proposition.

Proposition 4.2. *Suppose that the kernels $\phi(.,.), \psi(.,.)$ are symmetric and $\mathbb{P}_{YX} = \mathbb{P}_Y\mathbb{P}_X$, then the score-based kernel satisfies*

$$\mathbb{E}_{\boldsymbol{Z}_j\boldsymbol{Z}_q\boldsymbol{Z}_r}[\nabla_{\theta_k}\ell(\boldsymbol{z}, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)] = 0,$$
$$\forall 1 \leq k \leq d, \forall \boldsymbol{z} \in \mathbb{R}^{p+q},$$

so that the U-statistic based $\nabla_{\theta_k}\ell(\boldsymbol{z}, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)$ is degenerate.

Remark. The proof of this result can be found in the Supplementary material. As a consequence, we assume throughout this paper $\mathbb{P}_{YX} \neq \mathbb{P}_Y\mathbb{P}_X$, which implies that the variance of each $\nabla_{\theta_k}\ell(\boldsymbol{z}, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)$ is nonzero.

These variances form the diagonal of the variance covariance matrix $\mathbb{M}(\boldsymbol{\theta}_0)$. Each diagonal element corresponds to the variance:

$$\mathrm{Var}(\nabla_{\theta_k}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)), k = 1, \cdots, d,$$

which all are of degree 4, and each off-diagonal element corresponds to the covariance: $k, l \leq d, k \neq l$,

$$\mathrm{Cov}(\nabla_{\theta_k}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0), \nabla_{\theta_l}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)),$$

which are also kernels of degree 4. These quantities are defined in Eq. (7) and Eq. (8) in the Supplementary material.

In terms of population level, our parameter-dependent criterion is, for any θ, given by

$$\mathbb{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}_i\boldsymbol{Z}_j\boldsymbol{Z}_q\boldsymbol{Z}_r}[\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta})],$$

whose explicit expression in terms of the kernel is given in Eq. (14). In the rest of the paper, we will use the notation $\mathbb{E}[\ell(\boldsymbol{Z}; \boldsymbol{\theta})] := \mathbb{E}_{\boldsymbol{Z}_i\boldsymbol{Z}_j\boldsymbol{Z}_q\boldsymbol{Z}_r}[\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta})]$ when there is no confusion about the integration. The same applies to $\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z}; \boldsymbol{\theta})]$ and $\mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{Z}; \boldsymbol{\theta})\nabla_{\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z}; \boldsymbol{\theta})]$.

4.2 Asymptotic properties

In this section, we derive the large sample properties of the penalized estimator $\hat{\boldsymbol{\theta}}$ based on the M-estimation criterion:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Omega}{\mathrm{argmin}} \left\{ \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) + \sum_{k=1}^{d} \varphi\left(\frac{\lambda_n}{n}, |\theta_k|\right) \right\}, \tag{5}$$

with

$$\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) := (n)_4^{-1} \sum_{(i,j,q,r) \in \boldsymbol{I}_4^n} \ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}),$$

where $\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)$ is the symmetric kernel of degree 4 defined in Eq. (4). We assume that $\mathbb{L}(\theta) = \mathbb{E}[\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})]$ is uniquely minimized at $\boldsymbol{\theta}_0$. In addition, we make the following assumptions.

Assumption 1. *Sparsity assumption: $|\mathcal{A}| = k_0 < d$ with $\mathcal{A} = \{k : \theta_{0,k} \neq 0\}$.*

Assumption 2. *Distributive property of (Y_i, X_i): $\mathbb{P}_{YX} \neq \mathbb{P}_Y\mathbb{P}_X$.*

Assumption 3. *The parameter set Ω is a compact subset of $\mathbb{R}^d$.*

Assumption 4. *The kernels $\phi(.,.,), \psi(.,.)$ are symmetric and bounded.*

Assumption 5. *For any fixed $\boldsymbol{\theta}_0$, the matrices*

$$\mathbb{H}(\boldsymbol{\theta}_0) = \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z}; \boldsymbol{\theta}_0)],$$
$$\mathbb{M}(\boldsymbol{\theta}_0) = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{Z}; \boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z}; \boldsymbol{\theta}_0)],$$

exist and are positive definite.

Theorem 4.3. *Under assumptions 3-4, if $\frac{\lambda_n}{n} \to \lambda_0$, then for any compact $\boldsymbol{B} \subset \Theta$ such that $\boldsymbol{\theta}_0 \in \boldsymbol{B}$,*

$$\hat{\boldsymbol{\theta}} \underset{n \to \infty}{\xrightarrow{\mathbb{P}}} \underset{\boldsymbol{\theta} \in \boldsymbol{B}}{\mathrm{argmin}} \{\mathbb{L}^{pen}_\infty(\boldsymbol{\theta})\} = \boldsymbol{\theta}^*_0, \textit{ where}$$

$$\mathbb{L}_{\infty}^{pen}(\boldsymbol{\theta}) = \mathbb{L}(\boldsymbol{\theta}) + \sum_{k=1}^{d} \varphi(\lambda_0, |\theta_{0,k}|),$$

with $\mathbb{L}(\boldsymbol{\theta})$ given by Eq. (14) *evaluated at $\boldsymbol{\theta}$ and corresponding to the probability limit of $\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})$, and for any scalar θ, the penalty $\varphi(\lambda_0, |\theta|)$ is given as*

$$\begin{array}{lll}
\textbf{Lasso} & : & \lambda_0|\theta|, \\
\textbf{Bridge} & : & \lambda_0|\theta|^q, \\
\textbf{MCP} & : & \frac{b_{\text{mcp}}\lambda_0^2}{2}\mathbf{1}_{\{|\theta|>b_{\text{mcp}}\lambda_0\}} \\
 & & \quad -\frac{(b_{\text{mcp}}\lambda_0-|\theta|)^2}{(2b_{\text{mcp}})}\mathbf{1}_{\{|\theta|\leq b_{\text{mcp}}\lambda_0\}}, \\
\textbf{SCAD} & : & \begin{cases} \lambda_0|\theta|, & \text{for } |\theta| \leq \lambda_0, \\ \frac{-(\theta^2-2b_{\text{scad}}\lambda_0|\theta|+\lambda_0^2)}{(2(b_{\text{scad}}-1))}, & \text{for } \lambda_0 \leq |\theta| \leq b_{\text{scad}}\lambda_0, \\ (b_{\text{scad}}+1)\lambda_0^2/2, & \text{for } |\theta| > b_{\text{scad}}\lambda_0. \end{cases}
\end{array}$$

Hence if $\lambda_n = o(n)$, then $\hat{\boldsymbol{\theta}}$ is a consistent estimator.

Remark. The penalized estimator does not converge to $\boldsymbol{\theta}_0$ when $\lambda_n = O(n)$. In the first part of the proof, we prove the uniform convergence of the penalized criterion to the limit criterion. To do so, we use Theorem A.5 to derive a uniform law of large numbers of our parameter dependent U-statistic. We then rely on the convexity of the non-penalized criterion to deduce the consistency.

Assumption 6. *$\varphi(\frac{\lambda_n}{n}, |.|)$ is twice continuously differentiable except at the origin. We define*

$$A_{1,n} = \max_{k \in \mathcal{A}} |\nabla_{\theta_k} \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)|,$$

$$A_{2,n} = \max_{k \in \mathcal{A}} |\nabla^2_{\theta_k\theta_k} \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)|,$$

so that $A_{2,n} \to 0$.

Remark. The condition on the second derivative implies that the penalty has less influence than the non-penalized loss function in the regularized problem. Moreover, for the penalties of interest, the scaling of (λ_n, n) determines this rate.

Theorem 4.4. *Under assumptions 1-6, the sequence of penalized estimators $\hat{\boldsymbol{\theta}}$ satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2} + \sqrt{\text{card}(\mathcal{A})}A_{1,n}).$$

Remark. This result highlights that if $\lambda_n n^{-1} = O(n^{-1/2})$, then we would obtain a $\sqrt{n}$-consistent $\hat{\boldsymbol{\theta}}$. Note that the probability bound holds for any norm $\|.\|$. If we consider a setting with a diverging number of parameters – that is a double-asymptotic setting, where the dimension depends on the sample size – then the norm $\|.\|$ must be explicit because of the norm equivalences, where some constants may appear so that these constants may depend on the dimension.

We now derive the asymptotic distribution for the rate $\lambda_n = O(\sqrt{n})$ for Lasso, SCAD, and MCP and $\lambda_n = O(n^{q/2})$ in the Bridge case.

Theorem 4.5. *Under assumptions 1-6, suppose $\lambda_n = o(n)$; then if the regularization rates of Lasso, SCAD, and MCP satisfy $\lambda_n = O(\sqrt{n})$ and the Bridge regularization rate satisfies $\lambda_n = O(n^{q/2})$, and if $\lim_{\theta \to 0^+} \nabla_\theta \varphi(\frac{\lambda_n}{n}, \theta) = \frac{\lambda_n}{n}$ for SCAD and MCP, it follows that*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{d} \underset{\boldsymbol{u} \in \mathbb{R}^d}{\text{argmin}} \ \{\mathbb{F}_\infty(\boldsymbol{u})\},$$

provided $\mathbb{F}_\infty(.)$ is the random function in $\mathbb{R}^d$ where

$$\mathbb{F}_\infty(\boldsymbol{u}) = w^\top \boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^\top \mathbb{H}\boldsymbol{u} + \sum_{k=1}^{d} g(\lambda_0, u_k, \theta_{0,k}),$$

where $w \sim \mathcal{N}_{\mathbb{R}^d}(\mathbf{0}, \mathbb{M})$ with $\mathbb{M} := \mathbb{M}(\boldsymbol{\theta}_0)$, $\mathbb{H} := \mathbb{H}(\boldsymbol{\theta}_0)$ defined in assumption 5, and $g(\lambda, u, \theta)$ is given as follows:

$$\begin{array}{lll}
\textbf{Lasso} & : & \lambda\big(u \ \text{sgn}(\theta)\mathbf{1}_{\theta\neq 0} + |u|\mathbf{1}_{\theta=0}\big), \\
\textbf{Bridge} & : & \lambda|u|^q\mathbf{1}_{\theta=0} \\
\textbf{MCP} & : & \lambda|u|\mathbf{1}_{\theta=0}, \\
\textbf{SCAD} & : & \lambda|u|\mathbf{1}_{\theta=0}.
\end{array}$$

Remark. This result establishes the $\sqrt{n}$-consistency of the penalized estimator. However, for $\lambda_n = O(\sqrt{n})$ in the Lasso case, the term in $\mathbf{1}_{\theta_{0,k}\neq 0}$ implies that the true active set $\mathcal{A}$ cannot be recovered with high probability (see Proposition 1 of (Zou, 2006)).
To derive such distributions, we rely on specific theoretical results depending on the penalty function. In the Lasso case, because the objective function is convex, we are in a position to rely on the convexity Lemma B.1 of (Chernozhukov, 2005). For the non-convex MCP and SCAD cases, we rely on Lemma B.2 of (Umezu *et al.*, 2018). As for the Bridge, we lower-bound the asymptotic development by a quantity, whose minimum exists with high probability.

We now turn to the oracle property. It has been well known since (Zou, 2006) that Lasso does not satisfy this property. A way to fix this problem is to specify adaptive weights in the penalty function to penalize each coefficient differently. These adaptive weights are stochastic and depend on a first step estimator, which is required to be $\sqrt{n}$-consistent. In practice, this first step estimator is taken as a non-penalized M-estimator; that is, in the first step, the penalized criterion is solved for $\lambda_n = 0$. The key advantage of non-convex penalties is that they actually satisfy the oracle property without the need for these stochastic weights.

Theorem 4.6. *Suppose $\frac{\lambda_n}{n} \to 0$, for $\boldsymbol{\theta}_{\mathcal{A}}$ satisfying $\|\boldsymbol{\theta}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\| = O_p(n^{-1/2})$, under assumptions 1-6, suppose the MCP and SCAD regularization rates satisfy*

$\frac{\lambda_n}{n^{1/2}} \to \infty$ and

$$\lim_{n\to\infty}\inf \lim_{\theta\to 0^+}\inf \frac{n}{\lambda_n}\nabla_\theta \varphi\left(\frac{\lambda_n}{n}, \theta\right) > 0,$$

and suppose the Bridge satisfies the regularization rate $\frac{\lambda_n}{n^{q/2}} \to \infty, 0 < q < 1$ and $\lambda_n = O(\sqrt{n})$, then the $\sqrt{n}$-consistent local estimator $\hat{\boldsymbol{\theta}}$ defined in Theorem 4.4 satisfies $\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1$ and

$$\begin{aligned}&(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \theta_0) + \mathbf{S}_{n,\mathcal{A}\mathcal{A}})\sqrt{n}\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)_{\mathcal{A}}\\ &+ \left(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) + \mathbf{S}_{n,\mathcal{A}\mathcal{A}}\right)^{-1}\mathbf{b}_{n,\mathcal{A}}\}\\ &\xrightarrow[n\to\infty]{d} \mathcal{N}_{\mathbb{R}^{k_0}}(\mathbf{0}, \mathbb{M}),\end{aligned}$$

with $\mathbb{M} = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{Z}; \boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z}; \boldsymbol{\theta}_0)]$ and

$$\begin{aligned}\mathbf{b}_{n,\mathcal{A}} &= \big(\nabla_{\theta_1}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,1}|)\text{sgn}(\theta_{0,1}), \cdots,\\ &\qquad \nabla_{\theta_{k_0}}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k_0}|)\text{sgn}(\theta_{0,k_0})\big)^\top,\\ \mathbf{S}_{n,\mathcal{A}\mathcal{A}} &= \text{diag}(\nabla^2_{\theta_k\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|), k = 1, \cdots, k_0).\end{aligned}$$

Remark. This result establishes the conditions to satisfy the oracle property. Unlike Theorem 4.5, where the rate for SCAD/MCP is $\lambda_n = O(\sqrt{n})$ and that for the Bridge is $\lambda_n = O(n^{q/2})$, we now require $\lambda_n/\sqrt{n} \to \infty$ for the former cases and $\lambda_n/n^{q/2} \to \infty$ for the latter case. Note that adaptive Lasso (aLasso) is not reported because we discard the estimation methods requiring a two-step estimator as in the adaptive case (see (Zou, 2006)). The proof of the oracle property first focuses on the support recovery, that is, $\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1$. To do so, we prove that the sign of θ_k for indices $k \notin \mathcal{A}$ determines the sign of the score of the penalized criterion taken in θ_k under the assumption $\|\boldsymbol{\theta}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\| = O_p(n^{-1/2})$ for any given $\boldsymbol{\theta}$. We then derive the large sample distribution for the parameters whose indices belong to $\mathcal{A}$.

We propose to conclude our theoretical analysis with a consistency result when the dimension $d = d_n$ so that $d_n \to \infty$ when $n \to \infty$. The dimension satisfies $d_n = O(n^c)$ with some $q_1 < c < q_2$, $0 \leq q_1 < q_2 < 1$. We then have the following probabilistic bound.

Theorem 4.7. *Suppose that $d_n^2 = O(n)$, under assumptions 5 and 7-10, let $A_{1,n} = \max_{1\leq k\leq d_n} |\nabla_{\boldsymbol{\theta}_k}\varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)|$, then the sequence of penalized estimators $\hat{\boldsymbol{\theta}}$ satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O_p(\sqrt{d_n}(n^{-1/2} + \sqrt{\mathcal{A}_n}A_{1,n})).$$

Remark. Contrary to Theorem 4.4, where the bound holds for any norm, an explicit norm is required in the double asymptotic case. Indeed, because of the norm equivalences, some constants may appear so that these constant may depend on the size d_n and thus on n. All the assumptions adapted to the double asymptotic setting can be found in the Appendix.

5 Experiments

In this section, we illustrate the oracle property using a synthetic experience. We also carry out a real data experience to compare the forecasting performances.

5.1 Simulation experiment

We propose to explore the variable selection performance through the number of zero coefficients correctly estimated, denoted as C, the number of zero coefficients incorrectly estimated (i.e. an estimated zero coefficient whereas the true parameter is non-zero), denoted as $IC1$, the number of nonzero coefficients incorrectly estimated (i.e. an estimated non-zero coefficient whereas the true parameter is zero), denoted $IC2$, in Table 1, averaged for a hundred independent batches. Besides, the mean squared error is reported as an estimation accuracy measure. We consider the data generating process based on Eq. (3):

$$\forall i, j \leq n,\ \psi(Y_i, Y_j, \boldsymbol{H}_n) = \sum_{k=1}^{d}\theta_k\psi(X_i^{(k)}, X_j^{(k)}, \boldsymbol{H}_n) + U_{ij},$$

where the error term $U_{ij} \sim \mathcal{N}_{\mathbb{R}}(0, \zeta^2)$ with $\zeta = 2$ and $\forall k \leq d, \psi(X_i^{(k)}, X_j^{(k)}, \boldsymbol{H}_n) = (\bar{\boldsymbol{K}}^{(k)})_{ij}$ is evaluated over the multivariate Gaussian vector $X_o \sim \mathcal{N}_{\mathbb{R}^d}(\mathbf{0}, \Sigma), 1 \leq o \leq n$. In Eq. (3), we took $\phi(.,.) = \psi(.,.)$ and selected the Gaussian kernel:

$$\begin{aligned}\psi(Y_i, Y_j) &= \exp\left(-\tfrac{|Y_i - Y_j|^2}{2\sigma_y^2}\right),\\ \psi(X_i^{(k)}, X_j^{(k)}) &= \exp\left(-\tfrac{|X_i^{(k)} - X_j^{(k)}|^2}{2\sigma_x^2}\right),\end{aligned}$$

where σ_x^2 and σ_y^2 denote the widths of the kernel, which are set using the median heuristic (Sriperumbudur *et al.*, 2009): $\sigma_x = 2^{-1/2}\text{median}(\{\|X_i - X_j\|_2\}_{i,j=1}^n)$ and $\sigma_y = 2^{-1/2}\text{median}(\{\|Y_i - Y_j\|_2\}_{i,j=1}^n)$.

Note that our framework can accommodate alternative symmetric kernels such as the linear kernel, Laplace kernel, Abel kernel and the like. The variance covariance Σ is simulated such that $\Sigma_{pq} = \rho^{|p-q|}$ with $\rho = 0.8$. As for the true parameter vector, we consider $\forall i \leq k_0, \boldsymbol{\theta}_i \in \mathcal{U}([0, 2])$ the uniform distribution, whose true number of zero parameters k_0 depends on the problem size (and is arbitrarily set). We consider different problem sizes: $d = 400, 800$. In both cases, the true support is set as $k_0 = 100$. For the sample size, we considered $n = 1200$ when $d = 400$ and $n = 2000$ for $d = 800$. To recover the sparse support, we used Lasso, aLasso, SCAD, MCP, and Bridge.

To solve the penalization problem, in the SCAD and MCP cases, we apply a gradient descent method (see, *e.g.*, (Breheny and Huang, 2011)). For Lasso and its

Table 1: *Model selection and precision accuracy based on 100 replications.*

Model	MSE	C	IC1	IC2	MSE	C	IC1	IC2
Truth		300	0	0		700	0	0
Lasso	0.07	231.24	0.63	67.75	0.03	621.40	1.89	74.55
aLasso	0.59	300	16.25	0	0.08	700	13.75	0
SCAD	0.01	299.98	2.10	0.01	0.01	700	2.18	0
MCP	0.01	299.61	1.36	0.39	0.01	700	1.25	0
Bridge	0.07	300	3.64	0	0.05	700	3.22	0

Table 2: *Mean square error based on 100 test sets.*

Data Set	No Pen.	Lasso	aLasso	SCAD	MCP	Bridge
Isolet	2.085	1.965	1.847	1.843	1.829	1.837
Coil	3.250	3.037	2.949	2.951	2.960	3.014
ILPD	4.063	4.045	4.026	4.028	4.024	4.044

adaptive specification, we used the shooting algorithm developed by (Fu, 1998) and set the exponent entering the adaptive weights to $\gamma = 1.5$ (see equation 4 of (Zou, 2006)). We chose the non-penalized OLS estimator as the random coefficient entering these weights. Finally, we solved Bridge with $q = 1/2$ using a local quadratic approximation approach (see, e.g., (Fan and Li, 2001)). For selecting the regularization parameter, we used a standard cross-validation procedure.

Table 1 reports the performances of the regularization methods. The aLasso, SCAD, MCP, and Bridge perform better performances than Lasso in terms of variable selection and mean square error. This is in line with the asymptotic theory.

5.2 Real data experience

We carried out a performance analysis of the regularization methods on the real data sets Isolet, Coil, and ILPD from the UCI Machine Learning Repository. We considered a high-dimensional setting, where the number of observations is smaller than the number of covariates: for Isolet, we selected the first 150 observations and the first 500 covariates of the data set; for Coil, we selected the first 500 observations and the first 1000 covariates of the data set; we considered the full data set for the ILPD case, which is formed with 583 observations and 9 covariates.

The same V-statistic-based OLS problem as in the simulation experience is considered for prediction purposes with the Lasso, aLasso, SCAD, MCP, and Bridge regularization procedures. We also reported the non-penalized case (No Pen.). One hundred observations were then randomly chosen to fit the penalized OLS models, and the remaining 50 observations were used as a test set. The procedure was repeated 100 times, and the average mean square error for prediction is reported in Table 2.

The prediction performance is clearly improved when considering a penalized version of the regression model in the high-dimensional case that is, Isolet and Coil. Moreover, these results emphasize the advantage of using non-convex penalty functions. Although aLasso performs well, it still requires a $\sqrt{n}$-consistent first step estimator. The results are very close in the low-dimensional case.

6 Conclusion

In this paper, we proposed the SpHSIC regression, which is a versatile nonlinear feature selection algorithm. We obtained the conditions that satisfy the oracle property for the SpHSIC. Through experiments on synthetic and the real-world data, we demonstrated the ability of the proposed penalized model to recover the true sparse support.

In future work, the theoretical properties can be extended to a double-asymptotic setting, where the number of parameters diverges with the sample size. Some finite sample error bounds can also potentially be derived together with an evaluation of the probability that these bounds hold by using concentration inequalities.

Acknowledgments

M.Y. was supported by the JST PRESTO program JPMJPR165A and partly supported by MEXT KAKENHI 16H06299 and the RIKEN engineering network funding. B.P. was supported by the Start-Support Kakenhi 19K23193.

References

Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Computer Science*, **23**, 5–14.

Balasubramanian, K. *et al.* (2013). Ultrahigh dimensional feature screening via RKHS embeddings. In *AISTATS*.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied statistics*, **5**(1), 232.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, **33**(2), 806–839.

Chernozhukov, V. and Hong, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica*, **72**(5), 1445–1480.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition.

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

Fan, J. and 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.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, **3**(Mar), 1289–1305.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**(2), 109–135.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**(3), 397–416.

Gretton, A. *et al.* (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*.

Hastie, T. *et al.* (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Haws, D. C. *et al.* (2015). Variable-selection emerges on top in empirical comparison of whole-genome complex-trait prediction methods. *PloS one*, **10**(10), e0138903.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Statistical Research Report*, **5**(93).

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**(3), 293–325.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, **18**(1), 191–219.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*, **28**(5), 1356–1378.

Lee, A. (1990). *U-Statistics, Theory and Practice*. Dekker, New York.

Li, R. *et al.* (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**(499), 1129–1139.

Peng, H. *et al.* (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1237.

Ravikumar, P. *et al.* (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(5), 1009–1030.

Rejchel, W. (2017). Model selection consistency of u-statistics with convex loss and weighted lasso penalty. *Journal of Nonparametric Statistics*, **29**(4), 768791.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Song, L. *et al.* (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, **13**, 1393–1434.

Sriperumbudur, B. K. *et al.* (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS*.

Székely, G. J. *et al.* (2009). Brownian distance covariance. *The Annals of Applied Statistics*, **3**(4), 1236–1265.

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

Umezu, Y. *et al.* (2018). AIC for the non-concave penalized likelihood method. *Annals of the Institute of Statistical Mathematics*, **71**(2), 247–274.

Yamada, M. *et al.* (2014). High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, **26**(1), 185–207.

Yamada, M. *et al.* (2018). Ultra high-dimensional nonlinear feature selection for big biological data. *IEEE Transactions on Knowledge and Data Engineering*, **30**(7), 1352–1365.

Yeo, I.-K. and Johnson, R. A. (2001). A uniform strong law of large numbers for u-statistics with application to transforming to near symmetry. *Statistics & Probability Letters*, **51**(1), 63–69.

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

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