Feature Screening for High-dimensional Data in Classification Problems

Undergraduate Project

By

Shlok Mishra (210991)

BS Statistics and Data Science

Under the guidance of

Prof. Subhajit Dutta

Department of Mathematics and Statistics,

Indian Institute of Technology, Kanpur

November, 2023

Declaration

I/We hereby declare that the work presented in the project report entitled "Feature Screening for High-dimensional Data in Classification Problems" contains my own ideas in my own words. At places, where ideas and words are borrowed from other sources, proper references, as applicable, have been cited. To the best of our knowledge this work does not emanate from or resemble other work created by person(s) other than mentioned herein.

Date: Prof. Subhajit Dutta

Abstract

In this research, we introduce an innovative model-free feature screening methodology tailored to tackle ultrahigh-dimensional classification problems using univariate and bivariate Kolmogorov–Smirnov tests. This method stands out for its ability to effectively filter out irrelevant features and noise variables by examining both individual and pairwise variable distributions. In tandem with this screening technique, we have crafted a classifier based on kernel discriminant analysis, ensuring a harmonious integration between feature selection and classification mechanisms. Our comprehensive numerical experiments on simulated datasets highlight the robust performance of our approach, showcasing its empirical advantage in managing complex classification tasks.

Contents

1 Introduction

In a high-dimensional classification problem, the presence of a large number of irrelevant covariates (say, noise variables) usually deteriorates the performance of any classifier. Identifying relevant covariates (or, features), and subsequently, discarding noise often yield improved classification accuracy. Thus, feature screening continues to be an active area of research in classification problems. Existing screening methods are primarily comprised of two main steps:

- Rank the covariates according to their importance in predicting the response,
- Select the first $\tilde{n}$ of the ranked covariates, where $\tilde{n}$ is typically set to be $[n/\log n]$ and $[x]$ denotes the greatest integer less than or equal to $x \in \mathbb{R}$.

Let us illustrate using an example:

Example 1 *Let* $(X_1, \ldots, X_d) \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$ *and* $(Y_1, \ldots, Y_d) \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbf{I})$*, where* $\mathbf{0}$ *and* $\mathbf{I}$ *denote the null vector and the identity matrix, respectively, and* $\boldsymbol{\mu} = (1, 1, 1, 1, 0, \ldots, 0)^T$*. Here* $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *denotes the* d*-dimensional normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*.*

In this example, only the first four covariates are relevant for classification. Now, if we have 200 observations (i.e., $n = 200$), then all the aforementioned methods will select $\tilde{n} = 37$ covariates, irrespective of d. Eventually, the screened set will contain at least $37 - 4 = 33$ noise components. Clearly, the accumulation of noise in the screened set will have detrimental effects on the classification accuracy. The problem becomes even more severe when the sample size increases since $\tilde{n}$ is an increasing function of n.
A second major limitation of most of the existing screening methods is that they can only detect signals that arise from differences in marginal distributions, but are completely useless if the marginals are identical. We demonstrate this using a second example:

Example 2 *Suppose that* $(X_1, X_2), (X_3, X_4) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma_1})$ *with* $\boldsymbol{\Sigma_1} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ *and* $(Y_1, Y_2), (Y_3, Y_4) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma_2})$ *with* $\boldsymbol{\Sigma_2} = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$*, while* $X_5, \ldots, X_d$ *and* $Y_5, \ldots, Y_d$ *are iid* $\mathcal{N}(0, 1)$*. Here,* $(X_1, \ldots, X_4)$*,* $(X_5, \ldots, X_d)$*,* $(Y_1, \ldots, Y_4)$*, and* $(Y_5, \ldots, Y_d)$ *are mutually independent and 'iid' stands for independent and identically distributed.*

In this example, the pairs {1, 2} and {3, 4} contain signal through their bivariate distributions. But, the individual components are marginally undetectable since all the

one-dimensional marginals of the two competing distributions are $N(0,1)$. If we use any of the existing screening methods, it will select $\tilde{n}$ component variables (just like Example 1), which are all useless for classification.

After screening, our aim is to classify observations based on the screened features. In terms of compatibility, the criterion used for finding relevant features should also be reflected in the choice for the discriminant.

In this project, we propose a model-free screening method for ultrahigh-dimensional binary classification problems.

1. The proposed method retains only relevant features after eliminating noise variables, using a model-free approach that ensures the most significant predictors are selected for the model.

2. We have refined the method to identify pairs of features that are individually undetectable but exhibit significant interactions. Peacock's test, which extends the Kolmogorov–Smirnov test statistic to two-dimensional space, is employed to screen these paired features, ensuring that the two classes are differentiated with a maximum distinction in their joint distributions.

3. Lastly, our method stands out by maintaining coherence between the feature screening process and the subsequent classification rule. The classifier constructed is based on Kernel Density Estimation (KDE) and Kernel Discriminant Analysis (KDA), which aligns with the underlying principles of the feature screening technique.

2 Screening of Marginal Features

2.1 Kolmogorov–Smirnov (KS) Test on $\mathbb{R}$

The Kolmogorov–Smirnov statistic for a given cumulative distribution function (cdf) F is defined as:

$$\widehat{KS} = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|,$$

where $\sup_{-\infty<x<\infty}$ denotes the supremum of the set of distances, and $F_n(x)$ represents the empirical distribution function (edf). Given n independent and identically distributed (i.i.d.) observations $X_1, X_2, \ldots, X_n$ from the distribution function F, the edf is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

where I is the indicator function. Intuitively, the statistic captures the largest absolute difference between the two distribution functions over all possible values of x.

In the implementation of this project, the KS test is computed using the `ks.test()` function in R.

2.2 Univariate Two-Sample Kolmogorov–Smirnov Test Statistic

Given two independent samples $X_1, X_2, \ldots, X_{n_1}$ from distribution F and $Y_1, Y_2, \ldots, Y_{n_2}$ from distribution G, the two-sample Kolmogorov–Smirnov (KS) test statistic is defined as:

$$\widehat{KS} = \sup_{-\infty<x<\infty} |F_{n_1}(x) - G_{n_2}(x)|,$$

Where:

- $\sup_{-\infty<x<\infty}$ denotes the supremum of the set of distances across all values of x from $-\infty$ to $+\infty$.
- $F_{n_1}(x)$ is the empirical distribution function for the first sample X.
- $G_{n_2}(x)$ is the empirical distribution function for the second sample Y.

2.3 Methodology of Screening

Suppose that $\mathbf{F}$ and $\mathbf{G}$ are two absolutely continuous distribution functions (dfs) on $\mathbb{R}^d$. Let $\mathbf{X} = (X_1, \ldots, X_d)^\top \sim \mathbf{F}$ and $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top \sim \mathbf{G}$ with $X_k \sim F_k$ and $Y_k \sim G_k$ for $1 \leq k \leq d$. The covariate X_k can *marginally* discriminate between $\mathbf{F}$ and $\mathbf{G}$ if $F_k \neq G_k$. Consequently, the set of marginal signals is defined as

$$S = \{k : F_k \neq G_k \text{ for } 1 \leq k \leq d\}. \tag{1}$$

We denote its cardinality by $s := |S|$; the number of noise variables is $t := d - s$.

Now, we use the KS test statistic between the one-dimensional marginals of $\mathbf{F}$ and $\mathbf{G}$ to present an equivalent definition of S. The KS statistic between F_k and G_k is defined as

$$\mathcal{KS}_k = \sup_{-\infty<x<\infty} |F_k(x) - G_k(x)| \tag{2}$$

for $1 \leq k \leq d$. It is well known that $\mathcal{KS}_k = 0$ if and only if $F_k = G_k$ for $1 \leq k \leq d$.

Thus, KS statistics corresponding to the elements of S are strictly positive, whereas they are zero for the noise components. This allows us to express S as

$$S = \{k : \mathcal{KS}_k > 0 \text{ for } 1 \leq k \leq d\}, \tag{3}$$

and the screening problem reduces to identifying the covariates with positive KS statistic. Let us arrange the KS statistics $\{\mathcal{KS}_k : 1 \leq k \leq d\}$ in increasing order of magnitude.

Clearly, the smallest t KS statistics will correspond to the collection of noise variables and will all be equal to zero. In other words, we have

$$0 = \mathcal{KS}_{(1)} = \cdots = \mathcal{KS}_{(t)} < \mathcal{KS}_{(t+1)} \leq \cdots \leq \mathcal{KS}_{(d)}. \tag{4}$$

This now gives us yet another equivalent representation of S:

$$S = \{k : \mathcal{KS}_k \geq \mathcal{KS}_{(t+1)} \text{ for } 1 \leq k \leq d\}, \tag{5}$$

where $\mathcal{KS}_{(t+1)}$ denotes the minimum statistic among the signals, i.e., $\mathcal{KS}_{(t+1)} = \min_{k \in S} \mathcal{KS}_k$. A key observation here is that the ratio $\mathcal{KS}_{(t+1)}/\mathcal{KS}_{(t)} = \infty$, while $\mathcal{KS}_{(k+1)}/\mathcal{KS}_{(k)} < \infty$ for $(t+1) \leq k \leq (d-1)$.

In practice, both s and S are unknown and our objective is to estimate these quantities. We begin by estimating the KS test statistics. Assume $\min\{n_1, n_2\} \geq 2$. For random samples $\chi = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ and $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$ drawn from the dfs $\mathbf{F}$ and $\mathbf{G}$, respectively, an estimator of $\mathcal{KS}_k$ is given by the *sample KS statistic*:

$$\widehat{\mathcal{KS}} = \sup_x |\widehat{F}_{n_1}(x) - \widehat{G}_{n_2}(x)|, \tag{6}$$

for $1 \leq k \leq d$ where $\widehat{F}_{n_1}(x)$ and $\widehat{G}_{n_2}(x)$ are the empirical distribution functions for the k-th marginal of χ and $\mathcal{Y}$, respectively.

If we arrange the sample KS statistics in increasing order, following (4), one expects the first t values to correspond to elements of the noise set, while the complementary set should be the signal set. Using this idea, we first construct an estimate of S, the number of signals.

For absolutely continuous dfs F_k and G_k, we have $\widehat{\mathcal{KS}}_k > 0$ almost surely for $1 \leq k \leq d$. So, the ratio $\widehat{R}_k = \frac{\widehat{\mathcal{KS}}_{(k+1)}}{\widehat{\mathcal{KS}}_{(k)}}$ is well-defined with probability 1 for $1 \leq k \leq (d-1)$. Define R_k (the population counterpart of $\widehat{R}_k$) as $\frac{\mathcal{KS}_{(k+1)}}{\mathcal{KS}_{(k)}}$ for $t \leq k \leq (d-1)$. In view of the fact that $R_t = \frac{\mathcal{KS}_{(t+1)}}{\mathcal{KS}_{(t)}} = \infty$, we expect $\widehat{R}_t$ to take a significantly large value when compared with the entire sequence $\{\widehat{R}_k : 1 \leq k \leq (d-1)\}$. Detecting this jump yields the following estimator of the number of signals:

$$\widehat{t} = \arg \max_{1 \leq k \leq (d-1)} \widehat{R}_k, \quad \text{and} \quad \widehat{s} = d - \widehat{t}. \tag{7}$$

After obtaining $\widehat{s}$, we simply define the screened set to consist of the covariates corresponding to the $\widehat{s}$ largest $\widehat{\mathcal{KS}}_k$ values. In other words, we estimate the signal set S as:

$$\widehat{S} = \{k : \widehat{\mathcal{KS}}_k \geq \widehat{\mathcal{KS}}_{(\widehat{t}+1)} \text{ for } 1 \leq k \leq d\}. \tag{8}$$

Our proposed screening method is based on the idea of marginal differences, hence, we refer to it as *marginal screening* (MarKS).

3 Screening of Paired Features

3.1 Bivariate Kolmogorov-Smimov Test

3.1.1 Generalization to $\mathbb{R}^2$

The generalization of the Kolmogorov-Smirnov test to two-dimensional space involves more complexity. To generalize the test, Peacock [1983] proposed a procedure that involves four pairs of cumulative frequency functions. The samples in a plane are denoted by $\{(X_i^k, Y_i^k)\}_{i=1}^{n_k}$ for $k = 1, 2$. The cumulative frequency functions used by Peacock's test are given by

$$F_{++}^k(x,y) = \frac{\#\{i \mid X_i^k > x, Y_i^k > y, 1 \le i \le n_k\}}{n_k},$$
$$F_{+-}^k(x,y) = \frac{\#\{i \mid X_i^k > x, Y_i^k \le y, 1 \le i \le n_k\}}{n_k},$$
$$F_{-+}^k(x,y) = \frac{\#\{i \mid X_i^k \le x, Y_i^k > y, 1 \le i \le n_k\}}{n_k},$$
$$F_{--}^k(x,y) = \frac{\#\{i \mid X_i^k \le x, Y_i^k \le y, 1 \le i \le n_k\}}{n_k},$$

for $-\infty < x, y < \infty$ and $k = 1, 2$. Let $\{X_t^0 \mid t = 1, 2, \ldots, n\}$ be the pooled data set consisting of the values of the X-components of the given samples and $\{Y_t^0 \mid t = 1, 2, \ldots, n\}$ the pooled data set consisting of the values of the Y-components of the given samples. Define

$$D_{++} \stackrel{\text{def}}{=} \max_{1 \le s,t \le n} \left| F_{++}^1\left(X_s^0, Y_t^0\right) - F_{++}^2\left(X_s^0, Y_t^0\right) \right|,$$
$$D_{+-} \stackrel{\text{def}}{=} \max_{1 \le s,t \le n} \left| F_{+-}^1\left(X_s^0, Y_t^0\right) - F_{+-}^2\left(X_s^0, Y_t^0\right) \right|,$$
$$D_{-+} \stackrel{\text{def}}{=} \max_{1 \le s,t \le n} \left| F_{-+}^1\left(X_s^0, Y_t^0\right) - F_{-+}^2\left(X_s^0, Y_t^0\right) \right|,$$

and

$$D_{--} \stackrel{\text{def}}{=} \max_{1 \le s,t \le n} \left| F_{--}^1\left(X_s^0, Y_t^0\right) - F_{--}^2\left(X_s^0, Y_t^0\right) \right|.$$

Peacock's test is then defined as

$$D_{2DKS} = \max\{D_{++}, D_{+-}, D_{-+}, D_{--}\}$$

The test is often performed by a brute force algorithm and its application is very expensive in terms of computing time unless the sample sizes n_1 and n_2 are very small. Indeed, to compute the value of D_{--}, we need to compute the difference of the cumulative frequency

functions F^1_{--} and F^2_{--} at all the n^2 pairs (X_s, Y_t), with X_s and Y_t being coordinates of any pairs in the given samples. It will need $O(n)$ comparisons to compute the value of the difference at a single point. Thus, it will take $O(n^3)$ comparisons to compute D_{--}. Similar conclusions can be made for D_{++}, D_{+-}, and D_{-+}.

To alleviate the problem, Fasano and Franceschini [1987] revised Peacock's test by comparing the cumulative frequency functions at the observed sample points only, reducing the number of comparisons needed to $O(n^2)$. The F&F test is widely used in practice. However, it is a variant of Peacock's test and represents a different approach in essence. The complete algorithm is mentioned and has been discussed thoroughly in Lu et al. [2011]

In the implementation of this project, we utilized the 'peacock' R package, specifically employing the 'peacock2' function to perform the bivariate Kolmogorov-Smirnov tests. The 'peacock2' function implements the original definition of the two-dimensional Kolmogorov-Smirnov test as proposed by Peacock [1983], distinct from the Fasano and Franceschini [1987] test, which is a commonly used variant of the Peacock test.

The usage of 'peacock2' in our project is straightforward: 'peacock2(x, y)', where 'x' and 'y' represent the two samples being compared. Each sample is converted into a matrix where each row corresponds to a sample point, with the function using only the first two columns representing the X and Y components. The 'peacock2' test evaluates the test statistic over these matrices to determine the maximum discrepancy between the cumulative distribution functions of the two samples.

The Peacock test, as implemented by 'peacock2', has been integral to our analysis, allowing us to effectively compute the test statistic and thus contribute to the robustness of our feature screening process in the context of high-dimensional classification challenges.

3.2 Optimal Nonbipartite Matching

Matching serves as a cornerstone in statistical design and analysis, particularly for its ability to mitigate biases in observational studies. While bipartite matching has been the conventional approach, it is often restricted to simpler two-group designs. In contrast, nonbipartite matching can accommodate multiparty matching scenarios, making it suitable for designs with multiple treatment options.

Nonbipartite matching operates by finding the set of matches that minimize the sum of distances based on a provided distance matrix. This makes it particularly versatile for designs involving multigroup comparisons. Despite its potential, discussions and software implementations for nonbipartite matching are relatively sparse in the literature.

The algorithm seeks to:

- Provide a comprehensive solution for matching in scenarios with more than two treatment or group options.

- Handle complex designs by operating in a space that is not limited to bipartite structures, thus accommodating multiparty matching situations.
- Offer optimal matches that minimize the aggregate distance based on a given distance matrix, thus ensuring the best possible groupings.

The utility of this algorithm extends beyond simple matching, offering flexibility and broader applicability in various research designs. Its implementation has been made more accessible through dedicated software tools, ensuring that researchers can employ it effectively in their studies.

3.2.1 The Basic Idea of Matching and Optimal Matching

Matching methodology plays a fundamental role in observational studies, where the main objective is to compare outcomes of different treatment groups while controlling for confounding variables. For illustrative purposes, consider a simple design with one treatment group and one control group. In graph-theoretic terms, a graph consists of nodes and edges between nodes. Each individual unit of interest, such as a patient or hospital, is represented as a node. A matching is a collection of paired nodes between which an edge exists. Optimal matching aims to minimize the total distance among all pairs, which can be formulated as a restricted minimization problem. Various algorithms exist for this purpose, with the greedy matching algorithm being one of the most intuitive. However, this algorithm may not always yield the smallest total distance. The optimal matching that achieves this goal is referred to as an optimal matching. The matching that minimizes the total distance among all the pairs is called an optimal matching. It can be formulated as the solution to a restricted minimization problem. Without loss of generality, we focus on complete matching with an even number of nodes in the following discussion. Assume that every edge $[v_i, v_j] \in E$ is associated with a nonnegative weight w_{ij}. Then, as given in Xiao [2017], finding the optimal matching that minimizes the total distance is equivalent to finding a set of

$$x_{ij} = \begin{cases} 1 & \text{if } [v_i, v_j] \in M, \\ 0 & \text{if } [v_i, v_j] \notin M. \end{cases}$$

which solves the restricted minimization problem

$$\begin{aligned} \min \quad & \sum_{[v_i,v_j]\in E} w_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{j:[v_i,v_j]\in E} x_{ij} = 1 \quad \forall\, v_i \in V. \end{aligned}$$

3.2.2 Optimal Bipartite Matching and Optimal Nonbipartite Matching

Matching can be categorized based on the number of disjoint groups in a graph. Bipartite matching involves two disjoint groups, whereas nonbipartite matching can involve multiple groups. The former corresponds to the classical two-arm design in observational studies and has been extensively studied. Algorithms for optimal bipartite matching often frame the problem within a network flow framework. In contrast, optimal nonbipartite matching is more complex and requires different algorithms. One such algorithm involves searching for augmenting paths. The need for nonbipartite matching arises in studies with more intricate designs, where traditional two-group setups are insufficient.

3.3 Methodology of Screening

Let $\boldsymbol{F}_{\{i,j\}}$ and $\boldsymbol{G}_{\{i,j\}}$ denote the joint distributions of $\boldsymbol{X}_{\{i,j\}} = (X_i, X_j)^\top$ and $\boldsymbol{Y}_{\{i,j\}} = (Y_i, Y_j)^\top$, respectively, for $1 \leq i < j \leq d$. We define $\{i,j\}$ to be a *paired feature* if $F_i = G_i$ and $F_j = G_j$, but $\boldsymbol{F}_{\{i,j\}} \neq \boldsymbol{G}_{\{i,j\}}$. In other words, we have no discriminatory information in the marginal components $\{i\}$ and $\{j\}$, but only in the joint distribution through the pair $\{i,j\}$ for $1 \leq i < j \leq d$. The 2D KS test statistic between $\boldsymbol{F}_{\{i,j\}}$ and $\boldsymbol{G}_{\{i,j\}}$ using Peacock [1983]'s test is given by:

$$KS_{\{i,j\}} = \max\left\{D_{++_{\{i,j\}}}, D_{+-_{\{i,j\}}}, D_{-+_{\{i,j\}}}, D_{--_{\{i,j\}}}\right\},$$

where

$$\begin{aligned}
D_{++_{\{i,j\}}} &= \max_{1\leq s,t\leq n}\left|F_{++_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{++_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|,\\
D_{+-_{\{i,j\}}} &= \max_{1\leq s,t\leq n}\left|F_{+-_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{+-_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|,\\
D_{-+_{\{i,j\}}} &= \max_{1\leq s,t\leq n}\left|F_{-+_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{-+_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|,\\
D_{--_{\{i,j\}}} &= \max_{1\leq s,t\leq n}\left|F_{--_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{--_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|.
\end{aligned}$$

We have $\mathcal{KS}_{\{i,j\}} = 0$ iff $\mathbf{F}_{\{i,j\}} = \mathbf{G}_{\{i,j\}}$ for a pair $\{i,j\}$. As in Section 2, one may use $\mathcal{KS}_{\{i,j\}}$ to conclude whether a pair $\{i,j\}$ contributes to the signal, or not. In Example 2, the pairs $\{1,2\}$, $\{3,4\}$ are the only signals. So, $\mathcal{KS}_{\{1,2\}}$ and $\mathcal{KS}_{\{3,4\}}$ are positive, while $\mathcal{KS}_{\{i,j\}} = 0$ for all $\{i,j\} \neq \{1,2\}, \{3,4\}$. Using these facts from Example 2, we now develop the idea of screening paired signals.

We start by assuming d to be even. If not, we can make it even by adding an independently distributed noise term (e.g., a $N(0,1)$ variate). Let $\mathcal{P}_n$ denote the collection of all possible disjoint pairs which form a partition of $\{1,\ldots,d\}$, and define $\tilde{d} = d/2$. For a given partition $P = \left\{\{i_1, j_1\}, \ldots, \{i_{\tilde{d}}, j_{\tilde{d}}\}\right\} \in \mathcal{P}$, define $\mathcal{KS}(P) = \sum_{\{i,j\}\in P} \mathcal{KS}_{\{i,j\}}$. In Example 2, $\mathcal{KS}(P)$ can take four possible values, namely, $\mathcal{KS}_{\{1,2\}} + \mathcal{KS}_{\{3,4\}}$ if both $\{1,2\}$ and $\{3,4\} \in P_n$, $\mathcal{KS}_{\{1,2\}}$ if only $\{1,2\} \in P$, $\mathcal{KS}_{\{3,4\}}$ if only $\{3,4\} \in P$ and 0 otherwise.

Clearly, the maximum value that $\mathcal{KS}(P)$ can attain is $\mathcal{KS}_{\{1,2\}}+\mathcal{KS}_{\{3,4\}}$, and it is achieved when the partition P_n contains both the pairs $\{1,2\}$ and $\{3,4\}$. So, maximizing $\mathcal{KS}(P)$ over the set of all disjoint pairs $\mathcal{P}$ yields the set of paired signals. We now formalize this idea below.

Among the $\tilde{d}$ paired components, suppose that we have signal only in $s(<\tilde{d})$ paired features. In other words, let $i_1,\ldots,i_s$ and $j_1,\ldots,j_s$ be distinct integers in $\{1,\ldots,d\}$ such that $F_k = G_k$ for all $1 \le k \le d$, but $\mathbf{F}_{\{i_k,j_k\}} \neq \mathbf{G}_{\{i_k,j_k\}}$ for $1 \le k \le s$. This now implies that $\{i_1,j_1\},\ldots,\{i_s,j_s\}$ are the paired signals, while the rest of the components are noise. In this case, it clearly holds that

$$\mathcal{KS}(P) = \sum_{\{i,j\}\in P} \mathcal{KS}_{\{i,j\}} = \sum_{k=1}^{S} \mathcal{KS}_{\{i_k,j_k\}}\, \mathbb{I}[\{i_k,j_k\} \in P] \tag{9}$$

with $\mathbb{I}[\cdot]$ denoting the indicator function. The next result gives us a set of sufficient conditions under which $\mathcal{KS}(P)$ is maximized iff P contains all the paired signals.

Define

$$S = \big\{\{i_1,j_1\},\ldots,\{i_s,j_s\}\big\} \text{ and } S^c = \{1,\ldots,d\} \setminus \{i_1,\ldots,i_s,j_1,\ldots,j_s\}. \tag{10}$$

These can be viewed as the set of paired signals and noise components, respectively.

This formulation allows us to transform the problem of paired-feature screening into a maximization problem, with a nice interpretation from the graph-theoretic point of view. Let $G=(V,E)$ be an undirected graph with vertex set $V=\{1,\ldots,d\}$ and $\mathcal{KS}_{\{i,j\}}$ denote the weight of the edge between the i-th and j-th nodes for $1 \le i < j \le d$. Under this setting, maximizing $\mathcal{KS}(P)$ w.r.t P is equivalent to maximizing the sum of pairwise edge weights, where no two edges share the same node. This is the same as minimizing $-\sum_{\{i,j\}\in P} \mathcal{KS}_{\{i,j\}}$, or equivalently, $\sum_{\{i,j\}\in P}\big(M - \mathcal{KS}_{\{i,j\}}\big)$ for a constant $M > \max_{\{i,j\}\in P} \mathcal{KS}_{\{i,j\}}$. This essentially leads us to an optimal non-bipartite (NBP) matching problem of a graph with the weight of the (i,j)-th edge being $M - \mathcal{KS}_{\{i,j\}}$ for $1 \le i < j \le d$. Note that $\arg\max_{P\in\mathcal{P}} \mathcal{KS}(P)$ may not be unique.

Our goal is now to discard the noise pairs, and we adopt the same strategy as in the case of marginal signals. First of all, we define the empirical estimator of $\mathcal{KS}_{\{i,j\}}$ based on the training sample.

Given a set of observations $\{\boldsymbol{X}_k\}$ from $\boldsymbol{F}_{\{i,j\}}$ and $\{\boldsymbol{Y}_k\}$ from $\boldsymbol{G}_{\{i,j\}}$, we can define the

empirical cumulative distribution functions as

$$F_{++_{\{i,j\}}}(x,y) = \frac{1}{n_X}\sum_{k=1}^{n_X}\mathbb{I}\left(X_{i,k} > x, X_{j,k} > y\right),$$
$$F_{+-_{\{i,j\}}}(x,y) = \frac{1}{n_X}\sum_{k=1}^{n_X}\mathbb{I}\left(X_{i,k} > x, X_{j,k} \leq y\right),$$
$$F_{-+_{\{i,j\}}}(x,y) = \frac{1}{n_X}\sum_{k=1}^{n_X}\mathbb{I}\left(X_{i,k} \leq x, X_{j,k} > y\right),$$
$$F_{--_{\{i,j\}}}(x,y) = \frac{1}{n_X}\sum_{k=1}^{n_X}\mathbb{I}\left(X_{i,k} \leq x, X_{j,k} \leq y\right),$$

for $\boldsymbol{F}_{\{i,j\}}$, and similarly for $\boldsymbol{G}_{\{i,j\}}$:

$$G_{++_{\{i,j\}}}(x,y) = \frac{1}{n_Y}\sum_{k=1}^{n_Y}\mathbb{I}\left(Y_{i,k} > x, Y_{j,k} > y\right),$$
$$G_{+-_{\{i,j\}}}(x,y) = \frac{1}{n_Y}\sum_{k=1}^{n_Y}\mathbb{I}\left(Y_{i,k} > x, Y_{j,k} \leq y\right),$$
$$G_{-+_{\{i,j\}}}(x,y) = \frac{1}{n_Y}\sum_{k=1}^{n_Y}\mathbb{I}\left(Y_{i,k} \leq x, Y_{j,k} > y\right),$$
$$G_{--_{\{i,j\}}}(x,y) = \frac{1}{n_Y}\sum_{k=1}^{n_Y}\mathbb{I}\left(Y_{i,k} \leq x, Y_{j,k} \leq y\right),$$

where $\mathbb{I}$ is the indicator function, which is 1 if the condition is true and 0 otherwise.

The 2D KS test statistic between $\boldsymbol{F}_{\{i,j\}}$ and $\boldsymbol{G}_{\{i,j\}}$ using Peacock [1983]'s test is then given by

$$\widehat{\mathcal{KS}}_{\{i_k,j_k\}} = \max\left\{D_{++_{\{i,j\}}}, D_{+-_{\{i,j\}}}, D_{-+_{\{i,j\}}}, D_{--_{\{i,j\}}}\right\},$$

where

$$D_{++_{\{i,j\}}} = \max_{1\leq s\leq n_X, 1\leq t\leq n_Y}\left|F_{++_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{++_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|,$$
$$D_{+-_{\{i,j\}}} = \max_{1\leq s\leq n_X, 1\leq t\leq n_Y}\left|F_{+-_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{+-_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|,$$
$$D_{-+_{\{i,j\}}} = \max_{1\leq s\leq n_X, 1\leq t\leq n_Y}\left|F_{-+_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{-+_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|,$$
$$D_{--_{\{i,j\}}} = \max_{1\leq s\leq n_X, 1\leq t\leq n_Y}\left|F_{--_{\{i,j\}}}\left(X_s^0, Y_t^0\right) - G_{--_{\{i,j\}}}\left(X_s^0, Y_t^0\right)\right|.$$

Here, X_s^0 and Y_t^0 refer to the pooled samples from $\boldsymbol{X}$ and $\boldsymbol{Y}$.

We solve the optimal NBP matching by maximizing the following:

$$\widehat{\mathcal{KS}}(P) = \sum_{\{i,j\}\in P}\widehat{\mathcal{KS}}_{\{i,j\}} \text{ with respect to } P \in \mathcal{P}. \tag{11}$$

Let $\widehat{P} = \big\{\{i_1, j_1\}, \ldots, \{i_{\tilde{d}}, j_{\tilde{d}}\}\big\}$ denote a maximizer of (11). To reduce the notational burden, we denote $\widehat{\mathcal{KS}}_{\{i_k,j_k\}}$ simply by $\widehat{\mathcal{KS}}_k$ for $1 \leq k \leq \tilde{d}$ and denote their ordered values as $\widehat{\mathcal{KS}}_{(1)} \leq \widehat{\mathcal{KS}}_{(2)} \leq \cdots \leq \widehat{\mathcal{KS}}_{(\tilde{d})}$. Following the formulation of the MarKS algorithm (cf. (7) and (8)), we define

$$\widehat{t} = \underset{1\leq k\leq(\tilde{d}-1)}{\arg\max}\ \widehat{R}_k,\ \ \widehat{s} = \tilde{d} - \widehat{t} \text{ and } \widehat{S} = \big\{\{i_k, j_k\} : \widehat{\mathcal{KS}}_k \geq \widehat{\mathcal{KS}}_{(\widehat{t}+1)} \text{ for } 1 \leq k \leq \tilde{d}\big\}. \quad (12)$$

This screening method identifies differences between the joint distributions of pairs, hence, we refer to it as *paired screening* (or, PairKS).

Recall Example 2. Upon implementation, PairKS successfully screened both the pairs $\{1, 2\}$ and $\{3, 4\}$ in this example (more details in Results section). Moreover, PairKS did not select any noise component. The existing methods selected $\tilde{n} = 37$ components that contain no discriminatory information in their marginals and are clearly incapable of retaining the paired features.

4 Classifiers Incorporating Kernel Discriminant Analysis

Kernel Discriminant Analysis (KDA) is critical in enhancing classifiers, especially when dealing with non-linear and complex data structures. These methods are employed in classifiers using the screened set of features obtained from the feature screening process.
.

4.1 Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric technique used to estimate the probability density function of a continuous random variable. It is employed in a variety of applications, such as data visualization, statistical inference, and data modeling. The KDE can be thought of as a smoothed version of a histogram, where, rather than using bins to represent data frequencies, a kernel (a smooth, peaked function) is placed at each data point. The sum of these individual kernels then provides an estimate for the overall data distribution. Mathematically, the KDE is given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Where K is the kernel function, h is the bandwidth (which determines the width of the kernel), and n is the number of data points.

The kernel density estimate $\hat{f}_j$ is given by:

$$\hat{f}_j(\boldsymbol{x}; \mathbf{H}_j) = n_j^{-1} \sum_{i=1}^{n_j} K_{\mathbf{H}_j}(\boldsymbol{x} - \boldsymbol{X}_{ji}),$$

where $K_{\mathbf{H}_j}$ is a kernel function and $\mathbf{H}_j$ is a bandwidth matrix.

4.2 Kernel Discriminant Analysis

The main idea behind KDA is to use a kernel function to implicitly operate in a feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space.

By using kernel functions, KDA can capture complex, nonlinear relationships in the data and can result in better classification, especially when the classes are not linearly separable in the input space.

In practice, KDA involves determining optimal hyperplanes in the kernel-transformed space that best separate the different classes in the dataset. The choice of kernel function and its parameters play a critical role in the performance of KDA.

Moreover, the Kernel Discriminant Rule (KDR) is derived from the Bayes discriminant rule by incorporating kernel density estimates. Given ν populations, each associated with a density f_j and a prior probability π_j, $j = 1, 2, \ldots, \nu$, the Bayes discriminant rule allocates a sample point $\boldsymbol{x}$ to the group j_0 where:

$$j_0 = \underset{j \in \{1,2,\ldots,\nu\}}{\operatorname{argmax}} \pi_j f_j(\boldsymbol{x}).$$

The kernel discriminant rule is an approximation of this, replacing the true density f_j with its kernel density estimate $\hat{f}_j(\boldsymbol{x}; \mathbf{H}_j)$, which is computed from a random sample $\boldsymbol{X}_{j1}, \boldsymbol{X}_{j2}, \ldots, \boldsymbol{X}_{jn_j}$ representing the training data. The prior π_j is typically replaced by the sample proportion $\hat{\pi}_j = n_j/n$, where $n = \sum_{j=1}^{\nu} n_j$. The KDR then allocates $\boldsymbol{x}$ to the group j_0 as follows:

$$\text{KDR: Allocate } \boldsymbol{x} \text{ to group } j_0 \text{ ,where } j_0 = \underset{j \in \{1,2,\ldots,\nu\}}{\operatorname{argmax}} \hat{\pi}_j \hat{f}_j(\boldsymbol{x}; \mathbf{H}_j).$$

Kernel Discriminant Analysis (KDA) is critical in enhancing classifiers, especially when dealing with non-linear and complex data structures. These methods are employed in classifiers using the screened set of features obtained from the feature screening process.

4.3 Marginal Classifier with KDA

For the marginal classifier, KDA is used on the set of marginal signals S which have been identified as non-noise features. This screened set is defined as

$$S = \{k : F_k \neq G_k \text{ for } 1 \leq k \leq d\}. \tag{13}$$

where d is the total number of dimensions and the cardinality of S is denoted by s, with $s := |S|$ and the number of noise variables being $t := d - s$.

Given test data and the screened marginal variables in S, the discriminant function $DM(x)$ for the marginal classifier is computed as follows:

$$DM(\mathbf{x}) = \log \hat{f} - \log \hat{g},$$

where

$$\hat{f} = \frac{n_1}{n_1 + n_2} \prod_{k \in S} \hat{f}_k,$$

and

$$\hat{g} = \frac{n_2}{n_1 + n_2} \prod_{k \in S} \hat{g}_k.$$

Here, $\hat{f}_k$ and $\hat{g}_k$ are the KDA estimates for the respective classes over the screened marginals. The class label is determined based on the sign of $DM(x)$:

$$\delta_{KDA-MarKS} = \begin{cases} 1 & \text{if } DM(x) \geq 0, \\ 2 & \text{otherwise.} \end{cases}$$

4.4 Classifier for Paired Signals using KDA

For the classifier dealing with paired signals, the discriminant function $DP(x)$ is computed using only those pairs that belong to the screened set S. The class label assignment is then based on the value of $DP(x)$:

$$\delta_{KDA-PairKS} = \begin{cases} 1 & \text{if } DP(x) \geq 0, \\ 2 & \text{otherwise.} \end{cases}$$

In this approach, KDA facilitates the classification in the kernel-transformed feature space, enhancing the capability to discern complex patterns within the screened set S of paired signals.

Both classifiers utilize the strength of KDA to create more accurate models by focusing on the screened set S rather than the entire d-dimensional feature space, thereby improving performance in scenarios where traditional linear methods are inadequate.

5 Results

Our study evaluates the performance of the proposed screening algorithms, MarKS and PairKS, on simulations with 200 observations (100 for each class) and a high dimensionality of $d = 1000$. We refer back to Examples 1 and 2 discussed earlier and introduce four additional examples to illustrate the effectiveness of our methods.

In Examples 1–3 and 6, the noise variables are iid $N(0, 1)$. For Examples 4 and 5, the noise variables follow an iid $C(0, 1)$ distribution.

Example 3 $X_1, \ldots, X_4 \overset{iid}{\sim} N(0, 1)$ *and* $Y_1, \ldots, Y_4 \overset{iid}{\sim} N(0, 1/3)$.

Example 4 $X_1, \ldots, X_4 \overset{iid}{\sim} C(0, 1)$, *while* $Y_1, \ldots, Y_4 \overset{iid}{\sim} C(2, 1)$.

Example 5 $X_1, \ldots, X_4 \overset{iid}{\sim} C(0, 1)$ *and* $Y_1, \ldots, Y_4 \overset{iid}{\sim} C(0, 5)$.

Example 6 $X_1, \ldots, X_4 \overset{iid}{\sim} N(0, 4)$ *and* $Y_1, \ldots, Y_4$ *are iid from the mixture distribution* $\frac{1}{2}N(-\mu, 4 - \mu^2) + \frac{1}{2}N(\mu, 4 - \mu^2)$ *with* $\mu = 1.95$.

For each example, we computed the misclassification rates using classifiers adapted to the context: MarKS for all except Example 2, which uses PairKS. The following table compares these misclassification rates to the theoretical best rates achieved by a Bayes classifier.

Table 1: Misclassification rates for Examples 1-6

Example	Classifier	Bayes Rate (%)	Misclassification Rate (%)
1	MarKS	15.87	16.796
2	PairKS	5.07	5.394
3	MarKS	7.50	31.102
4	MarKS	11.32	23.560
5	MarKS	13.07	22.914
6	MarKS	5.09	4.032

The simulation setup involved imposing a restriction on the estimated signal set's cardinality, ensuring it did not exceed $[d/2]$. This restriction was due to the sparsity of features in our examples. Despite this, our proposed algorithms, MarKS and PairKS, demonstrated robust performance and can be applied even when the number of signals is relatively large.

The complete codebase for the methodologies and analyses discussed in this project is openly available for review and replication purposes. The relevant scripts can be accessed and downloaded from the following GitHub repository: shlokmishra/on-exact-feature-screening-ultrahigh-dimension.

6 Future Work

Before advancing to new screening methodologies, we must first address the remaining enhancements required for the PairKS screening method and its classifier. After these refinements are complete, our attention will shift to the development of the MixKS screening method, designed to adeptly manage mixed signals that include both paired and marginal features.

In the forthcoming months, my collaboration with Prof. Subhajit Dutta will be centered on perfecting the PairKS technique. Subsequently, we will embark on creating the MixKS approach. Our research to date has been concentrated on cases with either solely marginal or solely paired signals. The MixKS method aims to bridge this divide, offering a comprehensive solution for instances where signals are a confluence of both types.

Concurrent with the development of MixKS, we also plan to devise an efficient classifier tailored for the mixed signals scenario. Our ambition is to construct a robust system capable of integrating the nuances of mixed signal data within a cohesive analytical framework.

Additionally, our progression in the PairKS refinement and the genesis of the MixKS method will pave the way for our venture into multi-class classification problems. The extension of our techniques to accommodate J-class classification will signify a pivotal advancement, enhancing the versatility of our screening methods and classifiers to a wider spectrum of classification tasks.

References

J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 03 1983. ISSN 0035-8711. doi: 10.1093/mnras/202.3.615. URL https://doi.org/10.1093/mnras/202.3.615.

G. Fasano and A. Franceschini. A multidimensional version of the Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 03 1987. ISSN 0035-8711. doi: 10.1093/mnras/225.1.155. URL https://doi.org/10.1093/mnras/225.1.155.

Bo Lu, Robert Greevy, Xinyi Xu, and Cole Beck. Optimal nonbipartite matching and its statistical applications. *The American Statistician*, 65(1):21–30, 2011. URL https://EconPapers.repec.org/RePEc:bes:amstat:v:65:i:1:y:2011:p:21-30.

Yuanhui Xiao. A fast algorithm for two-dimensional kolmogorov–smirnov two sample tests. *Computational Statistics Data Analysis*, 105:53–58, 2017. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2016.07.014. URL https://www.sciencedirect.com/science/article/pii/S0167947316301785.