Feature Screening for High-dimensional Data in Classification Problems

Shlok Mishra

November 9, 2023

Introduction

- High-dimensional classification is often hindered by irrelevant features, leading to poor model performance.
- Effective feature screening is essential to enhance accuracy by isolating and removing noise.
- Existing methods rank features by predictive importance and select the top $\tilde{n}$ (where $\tilde{n}$ is set to $[n/\log n]$), potentially including many noise variables.
- Another major limitation of most of them is that they can only detect signals that arise from differences in marginal distributions, but are completely useless if the marginals are identical.

Proposed Method for Screening

- Our model-free approach aims for binary classification in ultrahigh-dimensional settings, emphasizing the preservation of relevant features.
- To screen marginal features, we implemented the Kolmogorov–Smirnov (KS) test on $\mathbb{R}$.
- For paired features, which are not detectable marginally but exhibit interaction, we employed Peacock's test—an extension of the KS test to two dimensions.
- Post-screening, our classifier, which is constructed for both marginal and pairwise features, employs Kernel Discriminant Analysis (KDA).

Motivation and Example

Example (Marginal Features)

Consider two multivariate normal distributions for classes:

- $(X_1, \ldots, X_d) \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$
- $(Y_1, \ldots, Y_d) \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbf{I})$

where $\mathbf{0}$ and $\mathbf{I}$ represent the null vector and identity matrix, respectively, and $\boldsymbol{\mu} = (1, 1, 1, 1, 0, \ldots, 0)^T$. In this setting, only the first four features are informative.

With $n = 200$, traditional methods would select $\tilde{n} = 37$ features, potentially including 33 noise variables, which could impair classification performance.

Motivation Continued with Another Example

Example (Paired Features)

Consider classes with pairs of features having strong interactions but identical marginals:

- $(X_1, X_2), (X_3, X_4) \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma_1})$
- $(Y_1, Y_2), (Y_3, Y_4) \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma_2})$

with $\mathbf{\Sigma_1} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ and $\mathbf{\Sigma_2} = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$, while other features are *iid* $\mathcal{N}(0, 1)$.

Here, signals are embedded in the covariance structure, undetectable through marginal distributions alone. Traditional methods would fail to identify these informative pairs, illustrating the need for our proposed screening approach.

Introduction to Marginal Screening

Marginal screening identifies covariates that individually discriminate between two distributions. It's based on the principle that if a covariate X_k has different distributions F_k and G_k in two groups, it can be informative for distinguishing between these groups.
We define the set of informative covariates as:

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

The Kolmogorov-Smirnov (KS) test statistic is used to measure the maximum difference between the empirical distribution functions of the covariates.

Estimating the KS Statistic

The KS statistic for a covariate X_k is defined as:

$$\mathcal{KS}_k = \sup_x |F_k(x) - G_k(x)|,$$

where F_k and G_k are the distribution functions for X_k under two different distributions.
For practical purposes, we use sample data to estimate the KS statistic:

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

where $\widehat{F}_{n_1}(x)$ and $\widehat{G}_{n_2}(x)$ are the empirical distribution functions based on the samples.

Screening Covariates

To screen covariates, we arrange the estimated KS statistics in ascending order. The covariates corresponding to the largest KS statistics are indicative of significant differences between the distributions, hence informative.
We estimate the number of signals (informative covariates) as:

$$\widehat{s} = d - \widehat{t},$$

where $\widehat{t}$ is estimated by the location of the largest jump in the ordered KS statistics, which differentiates noise from signals.
The screened set of covariates is then:

$$\widehat{S} = \{k : \widehat{\mathcal{KS}}_k \geq \widehat{\mathcal{KS}}_{(\widehat{t}+1)}\}.$$

This method is known as *marginal screening* (MarKS).

Marginal Variables Plot

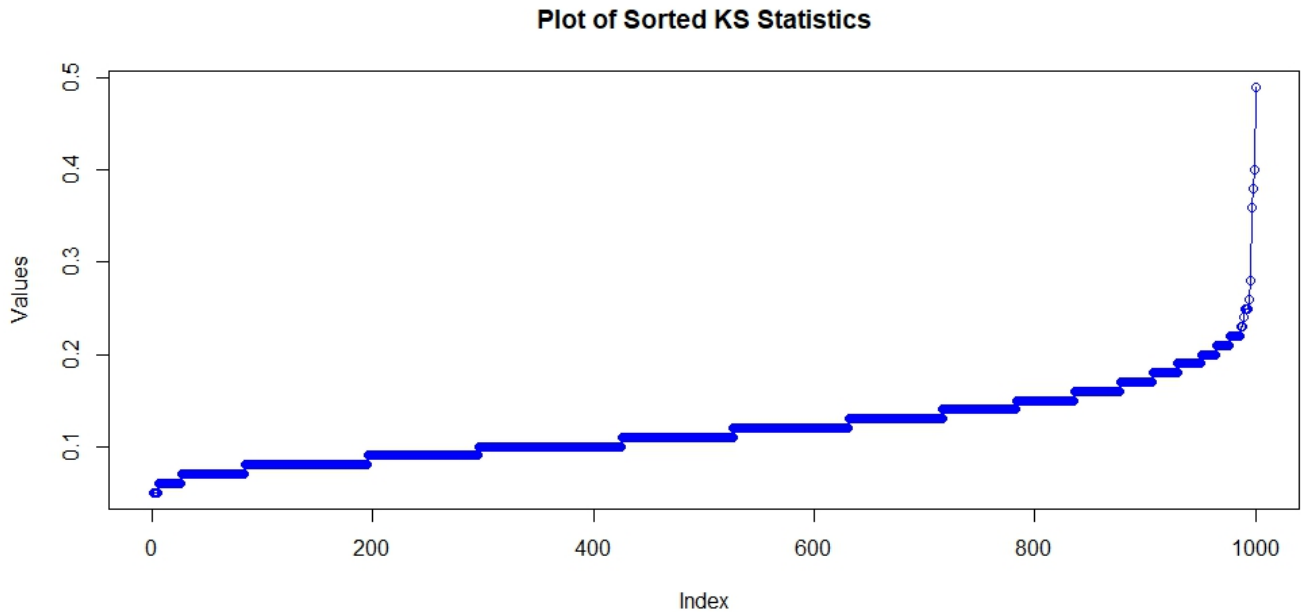

Figure: Clearly, the first four values of the KS statistic contribute to the signal, while the remaining 996 are considered noise.

Methodology of Pairwise Screening

Let $F_{\{i,j\}}$ and $G_{\{i,j\}}$ denote the joint distributions for paired features and define paired feature $\{i, j\}$ as one where $F_i = G_i$, $F_j = G_j$, but $F_{\{i,j\}} \neq G_{\{i,j\}}$. The 2D KS test statistic is given by:

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

with $D_{\pm\pm_{\{i,j\}}}$ being the maximum absolute difference in cumulative frequency functions.

Peacock's Test: Extension of KS Test on $\mathbb{R}^2$

- Peacock's test extends the KS test to 2D, considering cumulative distributions.
- It calculates the maximum discrepancy, D_{2DKS}, among four cumulative frequency function pairs.
- Let's focus on D_{--}, a component of D_{2DKS}, for an understanding of the test's intricacy.
- This test is computationally intensive, prompting the need for more efficient algorithms.

Peacock's Test: The D_{--} Component

- The D_{--} component measures the maximum difference where both variables in one sample are less than or equal to their counterparts in another.
- F^1_{--} and F^2_{--} represent the cumulative probability that a point from the first and second sample, respectively, falls into the lower-left quadrant defined by a point (x, y).
- Mathematically, D_{--} is defined as:

$$D_{--} \stackrel{\text{def}}{=} \max_{1 \leq s,t \leq n} \left| F^1_{--}\left(X^0_s, Y^0_t\right) - F^2_{--}\left(X^0_s, Y^0_t\right) \right|.$$

- This definition reflects the extremal difference in empirical distribution functions for the lower-left quadrant across two samples.

Example of Partitions and Signals

Given $d = \{1, \ldots, 6\}$, let us consider all possible disjoint pairs that form a partition of this set with $\tilde{d} = 3$. The pairs $\{1, 2\}$ and $\{3, 4\}$ are our signals.

The possible partitions P are:

- $P_1 = \big\{\{1, 2\}, \{3, 4\}, \{5, 6\}\big\}$ (contains both signals)
- $P_2 = \big\{\{1, 2\}, \{3, 5\}, \{4, 6\}\big\}$ (contains signal $\{1,2\}$)
- $P_3 = \big\{\{1, 2\}, \{3, 6\}, \{4, 5\}\big\}$ (contains signal $\{1,2\}$)
- $P_4 = \big\{\{1, 3\}, \{2, 4\}, \{5, 6\}\big\}$ (contains no signals)
- $P_5 = \big\{\{1, 3\}, \{2, 5\}, \{4, 6\}\big\}$ (contains no signals)
- $P_6 = \big\{\{1, 3\}, \{2, 6\}, \{4, 5\}\big\}$ (contains no signals)
- $\ldots$

Maximizing $\mathcal{KS}(P)$ over all partitions will identify the set of paired signals.

Graph-Theoretic Approach to Pairwise Screening

Define an undirected graph $G = (V, E)$ with vertices representing features and edges weighted by $\mathcal{KS}_{\{i,j\}}$ for the feature pairs. The objective is to maximize the sum of edge weights corresponding to disjoint feature pairs.

$$\mathcal{KS}(P) = \sum_{\{i,j\} \in P} \mathcal{KS}_{\{i,j\}}$$

This maximization is equivalent to solving an optimal non-bipartite (NBP) matching problem in the graph, with the modified edge weight $M - \mathcal{KS}_{\{i,j\}}$ for a large constant M. The set of paired signals is given by the matching that maximizes:

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

where $\mathcal{P}$ is the set of all possible partitions.

Identifying Significant Feature Pairs

Let $\widehat{P} = \{\{i_1, j_1\}, \ldots, \{i_{\tilde{d}}, j_{\tilde{d}}\}\}$ be the set of pairs maximizing the above. We simplify notation with $\widehat{\mathcal{KS}}_k$ for the k-th pair and order them as $\widehat{\mathcal{KS}}_{(1)} \leq \ldots \leq \widehat{\mathcal{KS}}_{(\tilde{d})}$.

The threshold $\widehat{t}$ is determined by the largest jump in ordered $\widehat{\mathcal{KS}}$ values, and the significant pairs are identified as:

$$\widehat{S} = \{\{i_k, j_k\} : \widehat{\mathcal{KS}}_k \geq \widehat{\mathcal{KS}}_{(\widehat{t}+1)}\}$$

This process, akin to the MarKS algorithm, is termed *paired screening* (PairKS), distinguishing joint distribution differences in feature pairs.

Marginal Classifier with KDA

KDA is applied to the marginal signals S to classify data points.
The set S consists of non-noise features where distributions F_k and G_k differ, for $1 \leq k \leq d$.
The discriminant function $DM(\mathbf{x})$ for a test point $\mathbf{x}$ is computed as the log ratio of the estimated class densities, using KDA over the screened marginals:

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

with $\delta_{KDA-MarKS}$ being 1 if $DM(x) \geq 0$ and 2 otherwise.

Classifier for Paired Signals using KDA

For paired signals, KDA is used to compute the discriminant function $DP(x)$, considering only those pairs in the screened set S. The resulting classifier $\delta_{KDA-PairKS}$ assigns class 1 if $DP(x) \geq 0$, and class 2 otherwise.
KDA enhances feature space analysis, allowing for complex pattern detection within the significant pairs, improving classification accuracy.

Simulation Study Overview

Our study evaluates the performance of MarKS and PairKS algorithms on simulations with:

- 200 observations (100 per class)
- A high dimensionality setting where $d = 1000$

We discuss the effectiveness of our methods through six examples, including:

- Examples 1–3 and 6 with noise variables as iid $N(0, 1)$
- Examples 4 and 5 with noise variables following an iid $C(0, 1)$ distribution

Example Specifications

Example

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

Example

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

Example

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

Example

$X_1, \ldots, X_4 \overset{\text{iid}}{\sim} N(0,4)$ and $Y_1, \ldots, Y_4$ follow a mixture distribution with $\mu = 1.95$.

Misclassification Rates

Table: 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	6.032

The methodologies and analyses for this project are available at: shlokmishra/on-exact-feature-screening-ultrahigh-dimension.

Optimal Nonbipartite Matching

- Nonbipartite matching is suitable for designs with multiple treatment options.
- It finds matches that minimize the sum of distances in a given distance matrix.
- This method is versatile and can handle complex designs.
- Matching is crucial in observational studies for comparing outcomes and controlling confounders.
- Optimal matching seeks to minimize the total distance based on a distance matrix.

Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric method to estimate the probability density function of a continuous random variable. It is useful in various domains such as data visualization and statistical inference. Unlike histograms, KDE places a kernel function at each data point and sums these to estimate the overall distribution.

The KDE for a set of data points is mathematically represented as:

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

Here:

- K is the kernel function—a smooth, peaked function.
- h is the bandwidth, which controls the kernel's width.
- n is the number of data points.

The estimate $\hat{f}_j$ with bandwidth matrix $\mathbf{H}_j$ is:

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

Kernel Discriminant Analysis (KDA) - Part I

Kernel Discriminant Analysis (KDA) leverages a kernel function to implicitly map data to a higher-dimensional feature space, enabling the operation on inner products without computing the actual coordinates.
The essence of KDA is to:

- Capture complex, nonlinear relationships within data.
- Improve classification when classes are not linearly separable in the original space.

It seeks optimal hyperplanes in the kernel-transformed space that best delineate the classes.
The kernel function's selection and parameters are crucial to KDA's success.

Kernel Discriminant Analysis (KDA) - Part II

The Kernel Discriminant Rule (KDR) is based on the Bayes rule, using kernel density estimates to classify a sample point $\boldsymbol{x}$:

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

Where:

- $\hat{\pi}_j$ is the sample proportion for the j-th group.
- $\hat{f}_j(\boldsymbol{x}; \mathbf{H}_j)$ is the kernel density estimate for $\boldsymbol{x}$ in the j-th group.

KDA is essential for classifiers to deal with non-linear data structures, utilizing the screened set of features.