On Exact Feature Screening in Ultrahigh-Dimensional Binary Classification

Sarbojit Roy, Soham Sarkar, Subhajit Dutta, and Anil K. Ghosh

Computer, Electrical, and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Delhi Centre, New Delhi, India; Department of Mathematics and Statistics, IIT Kanpur, Kanpur, India; Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India

ABSTRACT

We propose a new model-free feature screening method based on energy distances for ultrahigh-dimensional binary classification problems. With a high probability, the proposed method retains only relevant features after discarding all the noise variables. The proposed screening method is also extended to identify pairs of variables that are marginally undetectable but have differences in their joint distributions. Finally, we build a classifier that maintains coherence between the proposed feature selection criteria and discrimination method, and also establishes its risk consistency. An extensive numerical study with simulated and real benchmark datasets shows clear and convincing advantages of our proposed method compared to the state-of-the-art methods. Supplementary materials for this article are available online.

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 (see, e.g., Fan and Fan 2008; Pan, Wang, and Li 2016). In this article, we primarily focus on developing a screening method for binary classification problems that are free from model assumptions.

Existing screening methods are primarily comprised of two main steps. The first step is to rank the covariates according to their importance in predicting the response. A model-free feature screening method was developed by Zhu et al. (2011), where the conditional density of the response given a component variable was used to measure its importance. For binary classification problems, Mai and Zou (2013) introduced Kolmogorov filtering (referred to as KF) which uses the two-sample Kolmogorov-Smirnov statistic on the one-dimensional marginals to rank the covariates. A weighted average of the Cramer-von Mises distance between the conditional distribution function of a variable given the response and its unconditional distribution was proposed by Cui, Li, and Zhong (2015) to rank the covariates. Using this weighted average and the idea of sure independent screening (SIS) developed by Fan and Lv (2008), Cui, Li, and Zhong (2015) proposed the screening method MV-SIS. Cheng et al. (2017) developed a robust screening method (referred to as RRS) that uses the difference between the conditional rank of the covariates given the response and their unconditional ranks. VR-SIS was proposed by Song et al. (2018), where the authors used the ratio of the variance of a covariate conditioned on the response and the marginal variance as the marginal utility of the covariate. Using the idea of the Anderson-Darling test, the MV-SIS approach was modified in He, Ma, and Xu (2019). Covariates were ranked based on the classification accuracy of marginal classifiers (referred to as MCS) by Sheng and Wang (2020). Recently, Jiang et al. (2022) developed a new class of nonparametric test statistics, namely, the maximum adjusted Chi-squared (MAC) statistic for the two-sample testing problem, and proposed a variable screening procedure for binary classification problems using this MAC statistic. All these screening methods are model-free and possess the sure screening property (introduced by Fan and Lv 2008) in ultrahigh dimensions, viz., when the number of variables \( d = d_n \) satisfies \( \log d_n = O(n^\beta) \) for some \( \beta > 0 \) and \( n \in \mathbb{N} \). In other words, they retain all the features in the screened set with probability tending to one as the sample size tends to infinity, and the dimension is allowed to grow exponentially with the sample size.

After ranking the covariates, the second step is to select the first \( \tilde{n} \) of the ranked covariates, where \( \tilde{n} \) is typically set to be \( \lfloor n/\log n \rfloor \) (see Fan and Lv 2008). Here, \( \lfloor x \rfloor \) denotes the greatest integer less than, or equal to \( x \in \mathbb{R} \). Now, if the number of relevant features is strictly smaller than the number of selected components (i.e., \( \tilde{n} \)), then all these screening methods will inevitably include some noise components. We illustrate this using an example.

Example 1. Let \( (X_1, \ldots, X_{d_0}) \sim N_{d_0}(0, I) \) and \( (Y_1, \ldots, Y_{d_0}) \sim N_{d_0}(\mu, I) \), where \( 0 \) and \( I \) denote the null vector and the identity matrix, respectively, and \( \mu = (1, 1, 1, 1, 0, \ldots, 0) \). Here, \( N_{d_0}(\mu, \Sigma) \) denotes the \( d_0 \) dimensional normal distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \).
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.

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 \((X_1, X_2), (X_3, X_4) \sim N_2(0, \Sigma_1)\) with \(\Sigma_1 = [1, 0.9; 0.9, 1]\) and \((Y_1, Y_2), (Y_3, Y_4) \sim N_2(0, \Sigma_2)\) with \(\Sigma_2 = [1, -0.9; -0.9, 1]\), while \(X_5, \ldots, X_{d_1}\) and \(Y_5, \ldots, Y_{d_2}\) are independently 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 \( n \) components (just like Example 1), which are all useless for classification. There are some popular methods in the literature like group LASSO (see, e.g., Meier, Van De Geer, and Bühlmann 2008) and its sparse version (see, e.g., Simon et al. 2013) that can capture information from differences in joint distributions. More recently, Wang, Deng, and Xu (2023) proposed a grouped feature screening method based on the Gini impurity. However, these methods require apriori knowledge about plausible pairs and select the relevant ones out of those, and are not suitable in practice if the information on the pairs is unknown. To obtain paired features, Song et al. (2018) extended the VR-SIS method to select two covariates as a pair if their product yields a large value of their proposed VR index. Similar to their marginal screening method, the proposed modification also assumes the existence of second-order moments for the covariates, and hence, lacks robustness. Jiang et al. (2022) generalized their MAC based filtering method for screening features that are marginally undetectable, but have discriminatory information in their joint distributions. But, the performance of this method depends on a targeted false positive rate and this value needs to be tuned in practice.

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. For instance, suppose that the support vector machine with a radial basis kernel (say, SVMRBF) is used as the marginal classifier for selecting the top \( \tilde{n} \) variables. Since SVMRBF selected variables that contain discriminatory information between the competing populations in \( \mathbb{R}^{d_1} \), it will not be appropriate to use the SVM classifier with a linear kernel (say, SVMLIN) on the reduced space \( \mathbb{R}^{\tilde{n}} \). Among existing methods, only MCS possesses this congruity between the criteria for feature selection and discrimination.

Other methods only specify a screening procedure, but leave the choice of the classifier to the user.

In this article, we propose a model-free screening method for ultrahigh-dimensional binary classification problems. (a) With high probability, the proposed method retains only relevant features after discarding all the noises. (b) We extend the method to retain pairs along with marginal features. The pairs are screened in a way such that the two classes have maximum separation in terms of their energy distance. (c) Unlike almost all existing methods, there is a coherence between the methodology used for constructing the screening set and the proposed classification rule. It is worth pointing out that some existing methods satisfy a few of the above-mentioned properties (e.g., Fan and Song 2010; Hao, Feng, and Zhang 2018). However, such methods are model-based, whereas the primary focus of this article is to develop model-free screening methods coherent with the classification rule.

The rest of this article is organized as follows. Section 2 presents the proposed marginal screening method and theoretical results related to the consistency of the screened set. In Section 3, a further generalization of our method is developed that detects paired features by identifying differences between joint distributions of the bivariate components. Comparative performance of the existing and proposed algorithms is demonstrated using a variety of simulated datasets in Section 4. We propose a classifier that is coherent with the screening method in Section 5, and discuss related consistency properties. Our classification method is compared numerically with several popular classifiers on numerous simulated as well as real benchmark datasets in Sections 5.2 and 5.3, respectively. All proofs and mathematical details are provided in the Supplementary, which also contains algorithms of the proposed screening methods and some additional numerical results.

## 2. Marginal Screening Based on Energy Distances

Suppose that \( F \) and \( G \) are two absolutely continuous distribution functions (dfs) on \( \mathbb{R}^{d_1} \). Let \( X = (X_1, \ldots, X_{d_1})^\top \sim F \) and \( Y = (Y_1, \ldots, Y_{d_2})^\top \sim G \) with \( X_k \sim F_k \) and \( Y_k \sim G_k \) for \( 1 \leq k \leq d_n \).

The covariate \( X_k \) can marginally discriminate between \( F \) and \( G \) if \( F_k \neq G_k \). Consequently, the set of marginal signals is defined as

\[
S_n = \{k : F_k \neq G_k \text{ for } 1 \leq k \leq d_n\}.
\]

We denote its cardinality by \( s_n := |S_n| \); the number of noise variables is \( t_n := d_n - s_n \).

Now, we use energy distance between the one-dimensional marginals of \( F \) and \( G \) to present an equivalent definition of \( S_n \). Suppose that \( X_1, X_2 \) and \( Y_1, Y_2 \) are iid copies of \( X \) and \( Y \), respectively. Then, the energy distance between \( F_k \) and \( G_k \) is defined as

\[
\varepsilon_k = 2E[\gamma(|X_{1k} - Y_{1k}|^2)] - E[\gamma(|X_{1k} - X_{2k}|^2)]
- E[\gamma(|Y_{1k} - Y_{2k}|^2)]
\]

for \( 1 \leq k \leq d_n \). Here, \( \gamma \) is a continuous, monotonically increasing function from \( \mathbb{R}^+ \) to \( \mathbb{R}^+ \) such that \( \gamma \) has non-constant completely monotone derivative (Feller 1971) and \( \gamma(0) = 0 \). Some popular choices of the function \( \gamma(t) = 1 - \exp(-t) \).
After obtaining \( \hat{S}_n \), we simply define the screened set to consist of the covariates corresponding to the \( \hat{S}_n \) largest \( \hat{E}_k \) values. In other words, we estimate the signal set \( S \) as

\[
\hat{S}_n = \{ k : \hat{E}_k \geq \hat{E}_{(t_n+1)} \text{ for } 1 \leq k \leq d_n \}.
\] (8)

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

Note that energy based methods like the distance correlation (dCor) (see Székely, Rizzo, and Bakirov 2007) and the Hilbert-Schmidt Independence Criterion (HSIC) (see Gretton et al. 2005) have been used for developing variable screening methods in classification (see, e.g., Song et al. 2012; Balasubramanian, Sriperumbudur, and Lebanon 2013). These methods typically use tests of independence between the feature and the response to decide the relevance of a feature, whereas here, we use two-sample energy statistics to address this problem.

### 2.1. Screening Property of MarS

A screening method is said to possess sure screening property (SSP) if the estimated set \( \hat{S}_n \) contains the true signal set \( S \) with probability tending to one as the sample size increases, that is, \( P[S_n \subseteq S_n] \to 1 \) as \( n \to \infty \). The proposed method MarS possesses SSP under some regularity conditions. Recall that we work in the ultrahigh-dimensional regime, where \( \log d_n = O(n^\beta) \) for some \( 0 \leq \beta < 1 \). Consider the following assumptions:

- **A1.** There exists a constant \( 0 < \alpha_1 < (1 - \beta)/2 \) such that
  1. \( 1/\hat{E}_{(t_n+1)} = o(n^{\alpha_1}) \) and
  2. \( \max_{1 \leq k \leq d_n} R_k = o(n^{\alpha_1} \hat{E}_{(t_n+1)}) \).

Assumption A1.1 provides a lower bound on the rate of the minimum energy distance \( \hat{E}_{(t_n+1)} = \min_{k \in S_n} \hat{E}_k \) in the signal set \( S_n \). It is easy to see that the assumption can be restated as \( n^{\alpha_1}\hat{E}_{(t_n+1)} \to \infty \) as \( n \to \infty \) for some \( 0 < \alpha_1 < (1 - \beta)/2 \), which is equivalent to the existence of \( 0 < \tau < 1/2 \) such that \( n^{\tau}\hat{E}_{(t_n+1)} \to \infty \) is greater than some fixed constant. This is a common assumption in the variable screening literature and it readily holds if the energy distances between marginals are larger than some fixed constant. However, we also allow the minimum energy distance to go to zero, albeit at an appropriate rate. Assumption A1.2, on the other hand, states that the maximum value of \( R_k = \hat{E}_{(k+1)}/\hat{E}_k \) for \( (t_n + 1) \leq k \leq d_n \) cannot be too large compared to the minimum energy distance \( \hat{E}_{(t_n+1)} \). This is equivalent to requiring the difference between the successive energy distances in \( S_n \) to be controlled.

In other words, we put an upper bound on the rate of the relative growth of the energy distances in \( S_n \) (also see Section C.2 of the Supplementary). We now state the first theoretical result, which establishes SSP of the proposed method MarS.

**Theorem 1.** If assumption A1 is satisfied and \( \gamma \) is a bounded function, then there exists a constant \( b_1 > 0 \) (not depending on \( n \)) such that

\[
P[S_n \subseteq \hat{S}_n] \geq 1 - O \left( e^{-b_1 n^{1-2\alpha - \beta}} \right)
\]

for all \( 0 < \alpha_1 < \alpha < (1 - \beta)/2 \).

As a consequence, we have \( P[S_n \subseteq \hat{S}_n] \to 1 \) as \( n \to \infty \).
Among the popular choices of $\gamma$, the function $\gamma(t) = 1 - \exp(-t)$ for $t \geq 0$ is bounded.

**Remark 1.** The boundedness of $\gamma$ in the theorem is sufficient, but not necessary. For instance, if the univariate marginals of $F$ and $G$ are Gaussian and $\gamma$ is $L-$Lipschitz continuous for some $L > 0$ (e.g., $\gamma(t) = \log(1 + t)$ for $t \geq 0$ with $L = 1$), then Theorem 1 holds (see Lemmas V and VI in Appendix A for more details). Moreover, if the components of $X \sim F$ and $Y \sim G$ are bounded and $\gamma$ is continuous, then Theorem 1 remains valid.

Recall that we had used the ratios $\hat{R}_k$ for $1 \leq k \leq (d_n - 1)$ to construct the set $\hat{S}_n$. Now, if $\hat{R}_k$ attains the largest value for some $k < t_n$, then we are bound to include some noise variables in $\hat{S}_n$. Under an additional condition on the noise variables, we will prove that $\hat{S}_n$ not only retains all the signals but also disposes of all the noise components. Such a property is referred to as the exact screening property (ESP). To be precise, a screening method possesses ESP if the screened set $\hat{S}_n$ satisfies $P[\hat{S}_n = S_n] \to 1$ as $n \to \infty$. In other words, it perfectly estimates the true signal set with probability tending to one. Clearly, ESP is a stronger property than SSP. Now, consider the set $\mathcal{N}_n = \{\hat{R}_k : k \in S_n\}$ which is the collection of estimated energy distances corresponding to the noise variables. For notational ease, let $N_k$ denote the $k$th element and $N(k)$ be the $k$th minimum in the set $\mathcal{N}_n$ for $1 \leq k \leq t_n$. Now, consider the following assumption:

**A2.** There exists a constant $0 < \alpha_2 < (1 - \beta)/2$ such that

$$\max_{1 \leq k \leq (d_n - 1)} N_{k+1}/N_k = o_P(n^{\alpha_2}E_{(t_n+1)}).$$

**Theorem 2.** If assumptions A1 and A2 are satisfied for $0 < \alpha_1 < (1 - \beta)/2$ and $0 < \alpha_2 < (1 - \beta)/2$, respectively, and $\gamma$ is a bounded function, then there exists $b_2 > 0$ such that

$$P[\hat{S}_n = S_n] \geq 1 - O\left(e^{-b_2[n^{1-2\alpha}-n^{\alpha}]}ight)$$

for all $0 < \max\{\alpha_1, \alpha_2\} \leq \alpha < (1 - \beta)/2$.

As a consequence, $P[\hat{S}_n = S_n] \to 1$ as $n \to \infty$.

**Remark 2.** Theorem 2 can be proved under weaker versions of assumptions A1.2 and A2, with the “little $o$” being replaced by “big $O$” (see Remark II in the Supplementary for details).

Assumption A2 holds the key to proving ESP of MarS, and we first discuss this condition in details here. Fix $1 \leq k \leq t_n$. If $F_k = G_k$, then Székely and Rizzo (2005) showed that $n_1N_k/n$ converges in distribution to a nondegenerate random variable with a limiting (as $n \to \infty$) df (say, $H$) on $\mathbb{R}^+$. So, for large $n$, we can think of $n_1N_1/n, \ldots, n_1N_{t_n}/n$ as an ordered random sample of size $t_n$ from $H$. It is now clear that assumption A2 imposes a condition on the growth of the ratios of these ordered random variables, and allows these ratios to increase at a rate slower than $n^{\alpha_2}E_{(t_n+1)}$ for some $0 < \alpha_2 < (1 - \beta)/2$. By assumption A1.1, we already have $n^{\alpha_2}E_{(t_n+1)} \to \infty$ as $n \to \infty$. Using the results of Balakrishnan and Stepanov (2008) on the ratios of consecutive order statistics, one can show that $N_{k+1}/N_k = O_P(1)$ for any $1 \leq k \leq d_n$ (under appropriate regularity conditions on the df $H$). Further, if $L_n = \max_{1 \leq k \leq (t_n - 1)} N_{k+1}/N_k$ is $O_P(1)$, then A2 holds. In general, however, mathematical verification of this assumption is admittedly difficult.

We now check the validity of assumption A2 numerically. Let us consider a noise set $S_n$ with cardinality $t_n = [\exp(25n^{1/4})]$, where the components are iid $N(0,1)$. For a fixed $n$, let $N(i), \ldots, N(t_n)$ denote the ordered sample energy distances for the $i$th replicate with $1 \leq i \leq 100$. We compute $L_n(i) = \max_{1 \leq k \leq (t_n - 1)} N_{k+1}/N_k$ for $1 \leq i \leq 100$. The average of these values along with their standard errors are plotted against increasing $n$ in Figure 1. This figure clearly shows a decreasing trend as $n$ increases, indicating that $L_n$ is bounded in probability. We observe a similar phenomenon when the experiment was repeated with the standard Cauchy distribution (say, $C(0,1)$) as the noise distribution.

Let us now revisit Examples 1 and 2 introduced in Section 1. In Figure 2, we show the performance of different screening methods in these examples with $n = 200$ (100 observations from each class) and $d = 1000$ over 100 simulation runs. The
left panel of Figure 2 shows that MarS retained all the relevant features for most of the simulation runs. However, this was not the case for Example 2. The univariate marginals are all equal (viz., $F_k = G_k \equiv N(0, 1)$ for all $1 \leq k \leq d_n$) in this example. As a result, MarS as well as all competing methods fail to identify the signal set. To circumvent this problem, next, we propose a method that is capable of detecting differences in the joint distribution of pairs.

3. Screening of Paired Features

Let $F_{(i,j)}$ and $G_{(i,j)}$ denote the joint distributions of $X_{(i,j)} = (X_i, X_j)^T$ and $Y_{(i,j)} = (Y_i, Y_j)^T$, respectively, for $1 \leq i < j \leq d_n$. We define $\{i,j\}$ to be a paired feature if $F_i = G_i$ and $F_j = G_j$, but $F_{(i,j)} \neq G_{(i,j)}$. 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 $1 \leq i < j \leq d_n$. If $X_1, X_2$ and $Y_1, Y_2$ are iid copies of $X$ and $Y$, respectively, then the energy distance between $F_{(i,j)}$ and $G_{(i,j)}$ is given by

$$
E_{(i,j)} = 2E \left[ y \left( \frac{1}{2} \left\| X_{1,(i,j)} - Y_{1,(i,j)} \right\|^2 \right) \right] - E \left[ y \left( \frac{1}{2} \left\| X_{1,(i,j)} - X_{2,(i,j)} \right\|^2 \right) \right] - E \left[ y \left( \frac{1}{2} \left\| Y_{1,(i,j)} - Y_{2,(i,j)} \right\|^2 \right) \right] \quad (9)
$$

for $1 \leq i < j \leq d_n$. Here, $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^2$. We have $E_{(i,j)} = 0$ iff $F_{(i,j)} = G_{(i,j)}$ for a pair $\{i,j\}$. As in Section 2, one may use $E_{(i,j)}$ 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, $E_{(1,2)}$ and $E_{(3,4)}$ are positive, while $E_{(i,j)} = 0$ 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 $d_n$ 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 $P_n$ denote the collection of all possible disjoint pairs which form a partition of $\{1, \ldots, d_n\}$, and define $d_n = d_n/2$. For a given partition $P_n = \{\{i_1,j_1\}, \ldots, \{i_{d_n/2}, j_{d_n/2}\}\} \in \mathcal{P}_n$, define $\mathcal{E}(P_n) = \sum_{\{i,j\} \in P_n} E_{(i,j)}$. In Example 2, $\mathcal{E}(P_n)$ can take four possible values, namely, $\mathcal{E}(\{1\}) + \mathcal{E}(\{3,4\})$ if both $\{1,2\}$ and $\{3,4\}$ are in $P_n$, $\mathcal{E}(\{1\})$ if only $\{1,2\}$ is in $P_n$, $\mathcal{E}(\{3,4\})$ if only $\{3,4\}$ is in $P_n$ and 0 otherwise. Clearly, the maximum value that $\mathcal{E}(P_n)$ can attain is $\mathcal{E}(\{1\}) + \mathcal{E}(\{3,4\})$, and it is achieved when the partition $P_n$ contains both the pairs $\{1,2\}$ and $\{3,4\}$. So, maximizing $\mathcal{E}(P_n)$ over the set of all disjoint pairs $\mathcal{P}_n$ yields the set of paired signals. We now formalize this idea below.

Among the $d_n$ paired components, suppose that we have signal only in $s_n < d_n$ paired features. In other words, let $i_1, \ldots, i_{d_n}$ and $j_1, \ldots, j_n$ be distinct integers in $\{1, \ldots, d_n\}$ such that $F_i = G_i$ for all $1 \leq i \leq d_n$, but $F\{i,j\} \neq G\{i,j\}$ for $1 \leq i < j \leq s_n$. 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

$$
\mathcal{E}(P_n) = \sum_{\{i,j\} \in P_n} E_{(i,j)} = \sum_{k=1}^{s_n} \mathcal{E}(\{i_k,j_k\}) \quad \text{if } P_n \text{ contains all the paired signals.} \quad (10)
$$

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

Define

$$
S_n = \{i_1,j_1\}, \ldots, \{i_s,j_s\} \quad \text{and} \quad S_n^c = \{1, \ldots, d_n\} \setminus \{i_1, \ldots, i_s,j_1, \ldots, j_s\} \quad (11)
$$

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

**Lemma 1.** Suppose that $X_{(i,j)}$ and $X_{(i',j')} \in S_n$ are mutually independent for $\{i,j\}, \{i',j'\} \in S_n$ with $i \neq i'$ and $j \neq j'$. Also, let $X_{(i,j)}$ and $X_{(i',j')} \in S_n^c$ be mutually independent for any $\{i,j\} \in S_n^c$ and $i \in S_n^c$. Then, the maximum $\max_{P_n \in \mathcal{P}_n} \mathcal{E}(P_n) = \sum_{\{i,j\} \in S_n} \mathcal{E}(\{i,j\}) = \sum_{k=1}^{s_n} \mathcal{E}(\{i_k,j_k\})$ and the maximum is attained if $S_n \subseteq P_n$ for $P_n \in \mathcal{P}_n$. 

**Figure 2.** Bar plots indicating the average number of features selected by competing methods as well as the MarS algorithm in Examples 1 and 2 over 100 simulation runs with $n_1 = n_2 = 100$ and $d_n = 1000$. MarS was implemented with $\gamma(t) = 1 - \exp(-t)$ for $t \geq 0$. 

**Example 1**

| MCS | SVM-RBF | SVM-LIN |
|-----|---------|---------|
| KF  | RRS     | MV-SIS |

**Example 2**

| MCS | SVM-RBF | SVM-LIN |
|-----|---------|---------|
| KF  | RRS     | MV-SIS |

Size of the screened set

- Noise
- Marginal Features

0 4 10 20 30 37

0 4 10 20 30 37

\[ E_{\{i,j\}} = 2E \left[ y \left( \frac{1}{2} \left\| X_{1,(i,j)} - Y_{1,(i,j)} \right\|^2 \right) \right] - E \left[ y \left( \frac{1}{2} \left\| X_{1,(i,j)} - X_{2,(i,j)} \right\|^2 \right) \right] - E \left[ y \left( \frac{1}{2} \left\| Y_{1,(i,j)} - Y_{2,(i,j)} \right\|^2 \right) \right] \quad (9) \]
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 $V = \{1, \ldots, d_n\}$ and $E_{(i,j)}$ denote the weight of the edge between the $i$th and $j$th nodes for $1 \leq i < j \leq d_n$. Under this setting, maximizing $\mathcal{E}(P_n)$ w.r.t $P_n$ is equivalent to maximizing the sum of pairwise edge weights, where no two edges share the same node. This is the same as minimizing $-\sum_{(i,j) \in P_n} E_{(i,j)}$, or equivalently, $\sum_{(i,j) \in P_n} (M - \mathcal{E}_{(i,j)})$ for a constant $M = \max_{(i,j) \in P_n} \mathcal{E}_{(i,j)}$. This essentially leads us to an optimal non-bipartite (NBP) matching problem of a graph (see, e.g., Derigs 1988) with the weight of the $(i,j)$th edge being $(M - \mathcal{E}_{(i,j)})$ for $1 \leq i < j \leq d_n$. Note that $\arg\max_{P_n \in \mathcal{P}_n} \mathcal{E}(P_n)$ may not be unique. However, once we obtain a partition solving the NBP matching problem, Lemma 1 ensures that it contains all the signals and the remaining pairs are noise.

**Remark 3.** Although we prove Lemma 1 under the condition of independence of signal and noise variates, this is not necessary for the screening of paired features. In practice, we maximize the criterion (10), which gives us the pairs containing the maximum information for classification.

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 $\mathcal{E}_{(i,j)}$ based on the training sample as

$$
\hat{\mathcal{E}}_{(i,j)} = \frac{2}{n_1 n_2} \sum_{m_1=1}^{n_1} \sum_{m_2=1}^{n_1} \gamma \left( \frac{1}{2} \left\| X_{m_1,(i,j)} - Y_{m_2,(i,j)} \right\|^2 \right)
- \frac{1}{\binom{n_1}{2}} \sum_{1 \leq m_1 < m_2 \leq n_1} \gamma \left( \frac{1}{2} \left\| X_{m_1,(i,j)} - X_{m_2,(i,j)} \right\|^2 \right)
- \frac{1}{\binom{n_2}{2}} \sum_{1 \leq m_1 < m_2 \leq n_2} \gamma \left( \frac{1}{2} \left\| Y_{m_1,(i,j)} - Y_{m_2,(i,j)} \right\|^2 \right),
$$

(12) for $1 \leq i \neq j \leq d_n$. We solve the optimal NBP matching by maximizing the following:

$$
\hat{\mathcal{E}}(P_n) = \sum_{(i,j) \in P_n} \hat{\mathcal{E}}_{(i,j)} \text{ with respect to } P_n \in \mathcal{P}_n.
$$

Let $\hat{\mathcal{S}}_n = \{(i_1,j_1), \ldots, (i_{\tilde{d}_n},j_{\tilde{d}_n})\}$ denote a maximizer of (13). To reduce the notational burden, we denote $\hat{\mathcal{E}}_{(i_k,j_k)}$ simply by $\hat{\mathcal{E}}_k$ for $1 \leq k \leq \tilde{d}_n$ and denote their ordered values as $\hat{\mathcal{E}}_{(1)} \leq \hat{\mathcal{E}}_{(2)} \leq \cdots \leq \hat{\mathcal{E}}_{(\tilde{d}_n)}$. Following the formulation of the MarS algorithm (see, (7) and (8)), we define

$$
\tilde{\gamma}_n = \arg\max_{1 \leq k \leq (\tilde{d}_n-1)} \hat{R}_k, \quad \tilde{\gamma}_n = \tilde{d}_n - \tilde{\gamma}_n \quad \text{and} \quad \hat{S}_n = \{(i_k,j_k) : \hat{\mathcal{E}}_k \geq \hat{\mathcal{E}}_{(\tilde{d}_n+1)} \text{ for } 1 \leq k \leq \tilde{d}_n\}.
$$

(14)

This screening method identifies differences between the joint distributions of pairs, hence, we refer to it as paired screening (PairS). Based on our discussion above, it is clear that PairS will possess SSP/ESP under conditions similar to those used in Theorems 1 and 2.

Recall Example 2. The left panel in Figure 3 shows that PairS successfully screened both the pairs $[1,2]$ and $[3,4]$ in this example. Moreover, PairS 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. However, in the presence of a marginal feature, the NBP matching algorithm will forcefully couple the marginal component with some other (possibly a noise) component. We now introduce a third example to demonstrate this limitation of the PairS algorithm.

**Example 3.** $X_{[1,2]} \sim N_2(0, \Sigma_1)$ with $\Sigma_1 = [1, 0.9; 0.9, 1]$, $X_3 \sim N(1,1)$ and $Y_{[1,2]} \sim N_2(0, \Sigma_2)$ with $\Sigma_2 = [1, -0.9; -0.9, 1]$, while $X_4, \ldots, X_{d_n}$ and $Y_3, \ldots, Y_{d_n}$ are iid $N(0,1)$.

In Example 3, the pair $[1,2]$ contributes to the signal set through its joint (bivariate) distribution, while the third component contributes through the marginal distribution. The
remaining \((d_n - 3)\) components correspond to noise variables. The right panel of Figure 3 shows that in Example 3, competing methods successfully captured the marginal signal, but the remaining 36 components are all noises. The PairS method retained both \(1,2\) as well as \(3\), but the marginal signal \(3\) brought an additional noise with it. To summarize, neither the existing nor the proposed methods could retain both \(1,2\) and \(3\) as signals, and dispose of the remaining components as noise.

3.1. Screening of Mixed Features

The main issue with the PairS algorithm is that it always detects pairs of features. However, some of the screened pairs may contain a marginal feature and a noise, or two marginal features. Here we modify the PairS algorithm to circumvent this issue. Consider \(\hat{P}_n\), the partition obtained by solving the NBP matching problem. After screening significant pairs using the PairS algorithm, we obtain the estimated set \(\hat{S}_n\) given by (14). Without loss of generality, let us denote the screened pairs as \(\{(i_1,j_1)\ldots,(i_n,j_n)\}\). For each of these screened pairs, there are three possibilities: (i) one component is a marginal signal, while the other is a noise; (ii) both the components are marginal signals; and (iii) the individual components have no marginal signal, together they form a paired signal. More rigorously, for a screened pair \((i_k,j_k)\in \hat{S}_n\) for \(1 \leq k \leq n\), we have the following four possibilities:

1. \(i_k\) is a marginal signal, \(j_k\) is a noise, that is, \(F_{i_k} \neq G_{i_k}\) and \(F_{j_k} = G_{j_k}\),
2. \(j_k\) is a marginal signal, \(i_k\) is a noise, that is, \(F_{i_k} = G_{i_k}\) and \(F_{j_k} \neq G_{j_k}\),
3. both \(i_k\) and \(j_k\) are marginal signals, that is, \(F_{i_k} \neq G_{i_k}\) and \(F_{j_k} \neq G_{j_k}\),
4. \(i_k\) and \(j_k\) do not have marginal signals, but taken together they constitute a paired signal, that is, \(F_{i_k} = G_{i_k}\) and \(F_{j_k} = G_{j_k}\), but \(F_{(i_k,j_k)} \neq G_{(i_k,j_k)}\).

Among these four possibilities, we identify the most likely case among the following four mutually exhaustive null hypotheses is the most plausible:

\[
H_{1,0}^k : F_{i_k} = G_{i_k}, \quad H_{2,0}^k : F_{j_k} = G_{j_k},
H_{3,0}^k : F_{i_k} = G_{i_k}, F_{j_k} = G_{j_k} \text{ and } H_{4,0}^k : F_{(i_k,j_k)} = G_{(i_k,j_k)}.
\]

Using energy distances, these hypotheses can be equivalently expressed as follows:

\[
H_{1,0}^k : E_{i_k} = 0, \quad H_{2,0}^k : E_{j_k} = 0,
H_{3,0}^k : E_{i_k} + E_{j_k} = 0 \text{ and } H_{4,0}^k : E_{(i_k,j_k)} = 0.
\]

Although we have formulated the possibilities listed above as testing problems, it is not our aim to perform formal hypothesis testing. Instead, our goal is to check which of the four possible scenarios is the most plausible, which we achieve via comparing the corresponding \(p\)-values. In particular, we carry out the four hypotheses tests mentioned above using the energy distance-based testing procedure developed by Székely and Rizzo (2004). The test statistics of interest are \(\widehat{E}_{i_k}, \widehat{E}_{j_k}, \widehat{E}_{i_k} + \widehat{E}_{j_k}\) and \(\widehat{E}_{(i_k,j_k)}\), respectively, for \(1 \leq k \leq \hat{S}_n\). Let the corresponding \(p\)-values be \(p_1, p_2, p_3\), and \(p_4\), respectively. The minimum among these four \(p\)-values points toward the most likely null hypothesis to be rejected. For instance, if \(\min p_1 = p_1\), then we have the strongest evidence for \(F_{i_k} \neq G_{i_k}\), but \(F_{j_k} = G_{j_k}\). In this case, we retain \(i_k\) in the screened set as a singleton while discarding the \(j_k\)th component as a noise. The case \(\min p_1 = p_2\) is similar, and the signal and noise indices are essentially swapped. If \(\min p_1 = p_3\), then scenario 3 has the strongest evidence, that is, both \(i_k\) and \(j_k\) are screened as marginal signals. Finally, if \(\min p_1 = p_4\), then we retain \(\{i_k,j_k\}\) as a paired signal. We repeat this procedure for each of the selected pairs, and without loss of generality, we again denote the updated screened set as \(\hat{S}_n\). Since this modified method is capable of screening mixed features, that is, both marginal as well as paired features, we refer to it as mixed screening (MixS).

4. Numerical Studies

Using a diverse set of examples, we now study the performance of our proposed screening algorithms, viz., MarS and MixS (we exclude PairS due to its limitations and the subsequent introduction of MixS) and compare them with some existing methods. We have already discussed Examples 1–3 in earlier sections. Here, we introduce five new examples. In Examples 1–4 and 7–8, the noise variables are iid \(N(0,1)\). For Examples 5 and 6, the noise variables are iid \(C(0,1)\). The behavior of MarS and MixS is studied for the three different choices of the \(\gamma\) function mentioned after (2), viz., \(\gamma_1(t) = 1 - \exp(-t), \gamma_2(t) = (1+t)\) and \(\gamma_3(t) = \sqrt{t}\) for \(t \geq 0\).

**Example 4.** \(X_1, \ldots, X_4 \sim N(0,1)\) and \(Y_1, \ldots, Y_4 \sim N(0,1/3)\).

**Example 5.** \(X_1, \ldots, X_4 \sim C(0,1)\), while \(Y_1, \ldots, Y_4 \sim C(2,1)\).

**Example 6.** \(X_1, \ldots, X_4 \sim C(0,1)\) and \(Y_1, \ldots, Y_4 \sim C(0,5)\).

**Example 7.** \(X_1, \ldots, X_4 \sim N(0,4)\) and \(Y_1, \ldots, Y_4\) are iid from the mixture distribution \(0.5N(-\mu, 4 - \mu^2) + 0.5N(\mu, 4 - \mu^2)\) with \(\mu = 1.95\).

**Example 8.** \(X_1, \ldots, X_4 \sim N(0,1)\), while the pairs \(Y_{(1,2)}\) and \(Y_{(3,4)}\) are iid from the bivariate signed normal distribution (see Dutta and Genton 2014).

Example 4 is a scale problem with Gaussian distributions. Examples 5 and 6 correspond to a location and a scale problem, respectively, but involve heavy-tailed distributions. The parameters of the two distributions in Example 7 have been set in such a way that the means and variances are the same, but they have differences in their shapes. In Example 8, both \(Y_{(1,2)}\) and \(Y_{(3,4)}\) follow the bivariate signed normal distribution. The one-dimensional marginals are \(N(0,1)\) here, which implies that \(F_k = G_k\) for all \(1 \leq k \leq d_n\) in this example. Therefore, only \(\{1,2\}\) and \(\{3,4\}\) constitute the paired signals.

Throughout this study, we have simulated data with \(d_n = 1000\). In Examples 1–7, the training set is formed with 200 observations (100 from each class), while in Example 8 the training sample size is 400 (200 observations from each class).
All numerical results are based on 100 independent replications. The R package \texttt{VariableScreening} was used for implementing MV-SIS. We used the R package \texttt{e1071} for implementing \texttt{SVMLIN} and \texttt{SVMRBF} (related to MCS). For \texttt{SVMRBF} with kernel $K_{\theta}(x, y) = \exp(-\theta \|x - y\|^2)$, we considered the default value $1/d$ of the tuning parameter $\theta$. The implementation of NBP matching algorithm was done using the R package \texttt{nbpMatching} (see, e.g., Lu et al. 2011). Due to the sparsity of features in the considered examples, we impose a restriction on the estimated signal set so that its cardinality is less than, or equal to $[d_n/2]$ (see Appendix B in the Supplementary for more details). However, the proposed algorithms MarS and MixS can be readily used even when the number of signals is relatively large. Discarding half of the covariates might not be desired if the presence of a large number of signals is suspected. Under such circumstances, the proposed algorithms can be easily modified by discarding $[d_n/M]$ energy values for some large, positive value of $M$. One can also discard the lowest $d_{\text{min}}$ and the highest $d_{\text{max}}$ energies, where $d_{\text{min}}$ and $d_{\text{max}}$ are user-defined constants (see, e.g., Ni and Fang 2016).

The top left panel of Figure 4 shows that our proposed methods screen the signal components without selecting any additional noise in Example 1. We observe that the performance of MarS for $\gamma_1$ is a bit inferior when compared with $\gamma_2$ and $\gamma_3$. Since the univariate marginals are equal in Example 2, competing methods as well as MarS were totally useless here. However, the bivariate joint distribution had scale differences in the pairs $\{1, 2\}$ and $\{3, 4\}$, and the advantage of the MixS algorithm can be clearly seen from this example (see the top right panel of Figure 4).

The competing methods (except KF) managed to successfully screen the marginal features in Example 3, but at the cost of including 36 noise variables. In contrast, MarS selected the marginal features and discarded noise variables from $\hat{S}_n$. The MixS algorithm with $\gamma_1$ outperformed all other choices of $\gamma$ by effectively screening both $\{1, 2\}$ as well as $\{3\}$ (see the bottom left panel of Figure 4).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4}
\caption{Bar plots indicating the average number of variables selected by competing methods as well as the MarS and MixS algorithms in Examples 1–4 over 100 iterations.}
\end{figure}

Figure 4. Bar plots indicating the average number of variables selected by competing methods as well as the MarS and MixS algorithms in Examples 1–4 over 100 iterations.
panel of Figure 4). The bottom right panel of Figure 4 shows that the performance of both MarS and MixS is sensitive to the choice of $\gamma$ function in Example 4. In particular, $\gamma_1$ outperformed the other two choices here. Among the competing methods, RRS failed to retain any of the signals and detected 37 noise components, while the others managed to retain the marginal features (along with 33 noise variables).

Examples 5 and 6 involve heavy-tailed distributions. Observe from the top panel of Figure 5 that MCS-SVM with linear as well as RBF kernels were unsuccessful in both these examples. RRS too had poor performance in Example 6, which is a scale problem. Other competing methods screened all four signals but at the expense of gathering 33 noise components. In fact, none of the competing methods (except MV-SIS) achieved sure screening in these two examples. It is evident from the top panel of Figure 5 that $\gamma_3$ (an unbounded function) is a poor choice when dealing with heavy-tailed distributions in Examples 5 and 6. In contrast, $\gamma_1$ (a bounded function) yielded satisfactory results. However, MixS with $\gamma_1$ could not retain all four signals in Example 6. Recall that the screening accuracy of MixS relies on the solution of the NBP matching problem. Probably, the matching algorithm found a sub-optimal solution, which led to this deterioration in the overall performance. It is interesting to note that $\gamma_2$ outperformed $\gamma_1$ in both examples.

We dealt with a mixture distribution in Example 7. Among competing methods, only MCS-SVM (with the RBF kernel) and MV-SIS successfully retained all four relevant features (see the bottom left panel of Figure 5). MCS-SVMLIN could not perform well because the two classes do not have any difference in their mean vectors. KF picked only three important features, whereas the performance of RRS was worse as it failed to procure even a single signal. In stark contrast, MarS with $\gamma_1$ selected only the useful features and managed to discard noise, while MixS with $\gamma_1$ yielded quite encouraging results. Our proposals with $\gamma_2$ and $\gamma_3$ were less promising, but still performed better than the competing methods. Example 8 is more complex. Inescapably,
all the competing methods as well as the MarS algorithm failed to identify the paired signals. But, MixS successfully retained only the paired signals (see the bottom right panel of Figure 5). Among the three choices, $\gamma_1$ again led to the best result followed by $\gamma_2$ and $\gamma_3$. Appendix C.3 in the Supplementary contains a discussion on the relative performance of these three choices of $\gamma$.

Based on the above, we clearly see the usefulness of the proposed MarS and MixS algorithms. Among the competing methods, MV-SIS and MCS are generally quite promising in dealing with cases having marginal differences only. For the final task of classification, MV-SIS does not have a clear choice of classifier, which is left to the user. In contrast, MCS has a clear choice of the classifier (SVMLIN or SVMRBF). Hence, we decided to carry this method forward for a comparative performance in classification problems (see Section 5.2 for more details).

5. Classification Using Energy Distances

Our primary motivation behind using energy distances for screening was that they provide a measure of separation between the underlying distributions. As the next step, we use energy distances to build a classifier based on the signal set $S_n$. Recall that $S_n$ constitutes of two subsets, namely, $S_{1n}$ and $S_{2n}$ comprising of the marginal signals and the paired signals, respectively. Given $S_n$, we can define the average of energy distances between $F$ and $G$ as

$$\hat{E}_n = \frac{1}{s_{1n}} \sum_{k \in S_{1n}} \xi_k + \frac{1}{s_{2n}} \sum_{(i,j) \in S_{2n}} \xi_{(i,j)},$$

where $s_{1n} := |S_{1n}|$ and $s_{2n} := |S_{2n}|$. Fix $z \in \mathbb{R}^{d_n}$, and consider the following: $\xi^\gamma (z) = E[h_0^\gamma (z, Y_1)] - E[h_0^\gamma (z, X_1)] - \frac{1}{2} \{E[h_0^\gamma (Y_1, Y_2)] - E[h_0^\gamma (X_1, X_2)]\}$.

Here, $h_0^\gamma (u, v)$ is a measure of dissimilarity between $u, v \in \mathbb{R}^{d_n}$, which is defined as

$$h_0^\gamma (u, v) = \frac{1}{s_{1n}} \sum_{k \in S_{1n}} \gamma \left(|u_k| - |v_k|^2\right) + \frac{1}{s_{2n}} \sum_{(i,j) \in S_{2n}} \gamma \left(|u_i - v_j|^2 + |u_j - v_i|^2\right).$$

If $Z$ is independent of $X_1, X_2$ and $Y_1, Y_2$, then it is easy to show that $E_{Z \sim F} [\xi^\gamma (Z)] = \hat{E}_n$, while $E_{Z \sim G} [\xi^\gamma (Z)] = -\hat{E}_n$.

Since $\hat{E}_n$ is always positive, we expect $\xi^\gamma (z)$ to take positive (respectively, negative) values if $Z \sim F$ (respectively, $Z \sim G$). This motivates the use of $\xi^\gamma$ for classifying a test observation. Using $\xi^\gamma$, we define a classifier $\delta_0$ as follows:

$$\delta_0(z) = \begin{cases} 1, & \text{if } \xi^\gamma (z) > 0, \\ 2, & \text{otherwise}. \end{cases}$$

To use $\delta_0$ in practice, one needs to estimate the related quantities. Using the screened set $\hat{S}_n$ in place of $S_n$ (subsequently, $\hat{S}_{1n}$ and $\hat{S}_{2n}$ in places of $S_{1n}$ and $S_{2n}$, respectively) we obtain the sample counterpart of $h_0^\gamma (u, v)$ as

$$h_n^\gamma (u, v) = \frac{1}{s_{1n}} \sum_{k \in \hat{S}_{1n}} \gamma \left(|u_k| - |v_k|^2\right) + \frac{1}{s_{2n}} \sum_{(i,j) \in \hat{S}_{2n}} \gamma \left(|u_i - v_j|^2 + |u_j - v_i|^2\right).$$

Using $\hat{E}_n$, we define the sample version of $\delta_0$ as

$$\delta_{bgSAVG}(z) = \begin{cases} 1, & \text{if } \xi_n^\gamma (z) > 0, \\ 2, & \text{otherwise}. \end{cases}$$

This sample version is in fact the block generalized scale-adjusted average distance classifier (abbreviated as the bgSAVG classifier) proposed in Roy et al. (2022).

5.1. Consistency

Throughout this section, we assume that $S_n$ consists of marginal signals only, that is, $S_n = S_{1n}$ and $S_{2n} = \emptyset$ is the empty set. Consequently, $\hat{E}_n$ simplifies to $\sum_{k \in \hat{S}_{1n}} \xi_k / s_{1n}$. The definitions of $h_n^\gamma$ in (15) and $h_n^\gamma$ in (16) are also reduced to

$$h_n^\gamma (u, v) = \frac{1}{s_{1n}} \sum_{k \in \hat{S}_{1n}} \gamma \left(|u_k| - |v_k|^2\right)$$

respectively, for $u, v \in \mathbb{R}^{d_n}$. The resulting classifier is referred to as the generalized scale-adjusted average distance classifier (in short, gSAVG) classifier. We will study consistency properties of the gSAVG classifier combined with the marginal screening method MarS. Recall from Theorem 2 that if assumptions A1 and A2 are satisfied, then MarS possesses ESP. First, we state a convergence result for the discriminant $\xi_n^\gamma$.

Lemma 2. If assumptions A1 and A2 are satisfied for $0 < \alpha_1 < (1 - \beta)/2$ and $0 < \alpha_2 < (1 - \beta)/2$, respectively, and $\gamma$ is a bounded function, then there exists $b_3 > 0$ such that

$$P \left|\hat{E}_n^\gamma - \xi^\gamma (z)\right| > n^{-\alpha} \leq O \left(e^{-b_3(n^{-1 - \alpha} - n^{\alpha})}\right)$$

for all $z \in \mathbb{R}^{d_n}$ and $0 < \max(\alpha_1, \alpha_2) \leq \alpha < (1 - \beta)/2$.
From Lemma 2, it follows that $|\hat{\xi}_n (z) - \xi (z)|$ converges in probability to 0 at an exponential rate as $n \to \infty$.

The classifier $\delta_0$ is basically the "oracle" version of the classifier $\delta_{gSAVG}$. Let $\Delta_0$ and $\Delta_{gSAVG}$ denote the misclassification probabilities of $\delta_0$ and $\delta_{gSAVG}$, respectively. We now derive an upper bound for $\Delta_{gSAVG} - \Delta_0$, and analyze the terms in this bound as $n \to \infty$.

**Theorem 3.** If assumptions A1 and A2 are satisfied for $0 < \alpha_1 < 1 - \beta_2 / 2$ and $0 < \alpha_2 < 1 - \beta_2 / 2$, respectively, and $\gamma$ is bounded, then there exists a constant $b_3 > 0$ such that

$$\Delta_{gSAVG} - \Delta_0 \leq P[|\hat{\xi} (Z)| < n^{-\alpha}] + O\left(e^{-b_3(n^{1-2\alpha} - n^{\beta})}\right)$$

for all $0 < \max(\alpha_1, \alpha_2) \leq \alpha < (1 - \beta_2) / 2$.

The second term in the upper bound goes to zero as $n \to \infty$. Under some additional conditions, we can also show that $P[|\hat{\xi} (Z)| < n^{-\alpha}] \to 0$ and $\Delta_0 \to 0$ as $n \to \infty$ (see Lemmas VIII and IX in Appendix A of the Supplementary for details). This now readily implies that $\Delta_{gSAVG} \to 0$ (i.e., the gSAVG classifier achieves perfect classification) as $n \to \infty$. If $s_n \to \infty$, then this is expected because we accumulate more signal as $n \to \infty$, while ESP of MarS ensures that the noise variables get discarded. However, in our setup there is no restriction on the rate of growth of $s_n$, which can even be arbitrarily slow compared to $n$.

**Remark 4.** Alike Theorem 1, if we assume $F$ and $G$ to be Gaussian and $\gamma$ to be $L$-Lipschitz continuous for some $L > 0$ (e.g., $\gamma_2$), then Theorem 3 holds.

**Remark 5.** In Theorem 3, we have derived results for the case with marginal signals only. Similar derivations can also be made when we have paired, or mixed signals. While it is not difficult to make such an extension, the corresponding assumptions would be rather complicated. Hence, we chose not to pursue those to keep the exposition simple.

### 5.2. Comparison with Existing Classifiers

In this section, we conduct a comparative study of the performance of the proposed screening methods, namely, MarS and MixS in conjunction with the corresponding classifiers gSAVG and bgSAVG, respectively. We include some state-of-the-art classifiers for comparison. Recall that the MCS screening method with SVMLIN and SVMRBF classifiers were denoted by MCS-SVMLIN and MCS-SVMRBF, respectively. We shall continue to use the same notation for the respective classifiers. Apart from MCS, we also include some popular linear classifiers suitable for sparse data, viz., sparse Support Vector Machines (Sparse SVMLIN, Yi and Huang 2017), GLMNET (Hastie, Tibshirani, and Friedman 2009), sparse group LASSO (Sparse gLASSO, Simon et al. 2013) and the (non-linear) nearest neighbor classifier based on sparse random projections (Sparse NN-RAND, Deegalla and Bostrom 2006).

The R packages sparsesVM, glmnet, SGL and RcmdPro were used for implementation of Sparse SVMLIN, GLMNET, Sparse gLASSO and Sparse NN-RAND, respectively. In this study, the parameters of the above methods were set to default (unless specified otherwise). For Sparse SVMLIN, the parameter associated with the penalty function was tuned using cross-validation. In GLMNET, the tuning parameter $\lambda$ was chosen by cross-validation (with the number of folds set to be the default value 10). We provided the group memberships during our implementation of Sparse gLASSO, while the penalty term was chosen using cross-validation. For Sparse NN-RAND, we used the projection method "li" that generates very sparse random matrices. R codes for the proposed classifiers are available at [https://github.com/RSarbojot/energy_screening_ultrahigh_dimensional](https://github.com/RSarbojot/energy_screening_ultrahigh_dimensional).

We first implemented the gSAVG classifier without any screening (WoS). Then, the gSAVG classifier was combined with MarS, while the bgSAVG classifier was used with MixS. Recall Examples 1–8 from Section 4. The training sample in each of Examples 1–7 was formed by generating 100 observations from each class and a test set of size 500 (250 from each class) was used. In Example 8, the training sample constituted of 200 observations from each class, while the test size remained 500. This procedure was repeated 100 times to compute the average misclassification rates, which are reported in Table 1.

Superiority of the gSAVG classifier with MarS is clear from the examples with marginal differences, namely, Examples 1, 4, 5, 6, and 7 (see Table 1). On the other hand, the bgSAVG classifier with MixS yielded promising results in the non-marginal scenarios (namely, Examples 2, 3, and 8). Among the competing methods, linear classifiers showed satisfactory performance only in Examples 1 and 3. They failed in Examples 2, 4, 7, and 8 where the difference lies either in scale or shape, and led to a misclassification rate of almost 50%. The performance of Sparse NN-RAND was generally quite poor across all examples. Only MCS-SVMRBF showed some improvement in location and scale problems. In Examples 5 and 6 (with heavy-tailed distributions), none of the competing methods yielded satisfactory results due to lack of robustness.

We have also conducted a thorough numerical study to investigate the sensitivity of the proposed screening methods toward our assumption of independence between the signal and noise variables. To summarize our findings, it is clear that the proposed methods perform well even under departure from this assumption (see Appendix D.1 of the Supplementary for details). Our theoretical results further assume sparsity in the feature variables. We have investigated the empirical performance of both the MarS and MixS algorithms when the number of signals increases, while the dimension is kept fixed. It is transparent from our analyses in Appendix D.2 of the Supplementary that the proposed methods can be conveniently used even when the sparsity assumption is violated.

### 5.3. Real Data Analysis

We now study the performance of our proposed methodology on two benchmark and four real datasets. The first data is Madelon, which is an artificial dataset that was part of the NIPS 2003 feature selection challenge (Guyon et al. 2005). The data points are of 500 dimension with 5 informative covariates. An additional 15 linear combinations of these covariates were added to form a signal set having cardinality 20. The remaining 480 components are all noise with no predictive
power. Madelon has a total of 2600 observations (1300 from each class). The second dataset CorrAL (John, Kohavi, and Pfleger 1994) is also an artificial data of dimension 10,000. It has 128 observations from two classes, 72 from the first class and 56 from the second class. Out of the 10,000 features, only the first 6 components are highly correlated with the response variable, while the rest are all noise. We used a subset of this dataset containing the first 1000 features. Both Madelon and CorrAL are available from the R package sbfe. The third and fourth datasets, Bittner and Shipper are from the Compcancer database. Bittner is a microarray dataset with 38 observations of dimension 2201. This dataset was first studied by Bittner et al. (2000), where the authors analyzed gene expression profiles for the 38 samples. Class labels of 19 tightly clustered samples was denoted by ML1, while the remaining 19 samples were labeled as ML2. Shipp was introduced by Shipp et al. (2002), where the authors aimed at distinguishing diffuse large B-cell lymphoma (DLBCL) from a related GC B-cell lymphoma, follicular (FL). It has 77 data points (58 and 19 samples corresponding to DLBCL and FL, respectively) of dimension 798. The final two datasets are related to Microarray gene expressions. The GSE726 data has 52 gene expressions of dimension 2283 from breast and colon cancer patients with 31 and 21 samples, respectively. In the GSE967 dataset, we have 23 observations of dimension 9945. These gene expression values correspond to two childhood tumors, namely, Ewing’s sarcoma (EWS) and rhabdomyosarcoma (RMS) containing 11 and 12 data points.

For each of these six datasets, we randomly selected 50% of the observations (without replacement) corresponding to each class to form the training set. The remaining observations were considered as test cases. This procedure was repeated 100 times over different splits of the data to obtain a stable estimate of the misclassification rate. As already noted in our simulations, the number of features selected by MarS and MixS varies with the choice of γ. For each real dataset, we have summarized this information in the respective paragraphs, while the complete numerical result is presented in Appendix C of the Supplementary (see Table 1).

We start by analyzing the Madelon data. The number of components screened by MCS was 181. Even if all 5 signals are retained by MCS, its screened set still contained about 97% noise. In contrast, MarS selected 3.31 components, while the number of features screened by MixS ranged between 7.78 and 8.39 on an average for the three choices of γ. Usefulness of feature screening is clear from this dataset (see Table 2). The gSAVG classifier without any screening (WoS) performed worse than gSAVG with MarS and bgSAVG with MixS. In particular, MixS for γ3 when combined with the bgSAVG classifier yielded the minimum misclassification rate. Among the competing classifiers, only GLMNET led to a competitive performance.

MCS screened 15 components in the CorrAL data. MarS screened 1.54 components, while MixS selected 2.46 components on an average for varying γ. Again, we observe a significant reduction in misclassification rate of the proposed screening methods w.r.t. the WoS version of the gSAVG classifier. Among the competing methods, MCS-SVM with the RBF kernel produced the best result. While GLMNET successfully detected sparsity and its performance was at par with the proposed methods, the others did not fare so well.

In Bittner, MCS screened about 6 components out of 2201. On an average, the number of components retained by MarS ranged from 3.51 to 3.6, whereas for MixS the range was from 4.9 to 5.6 for the three choices of γ. Interestingly, Table 2 shows that the performance of gSAVG and bgSAVG is subjective to the choice of the γ function. Overall, bgSAVG with MixS for γ3 led to the minimum classification rate, while gSAVG with MarS for γ2 secured the second position. MCS-SVMLIN led to the lowest misclassification rate among the competing methods, closely followed by GLMNET and Sparse NN-RAND.

For different choices of γ, the average number of variables screened by MarS in the Shipp data ranged from 2.52 to 4.75, while it ranged from 3.66 to 5.20 for MixS. Superiority of MixS over MarS is transparent from this dataset (see Table 2). In particular, bgSAVG with MixS had a significantly improved performance when compared to gSAVG with MarS for the choices γ2 and γ3. The performances of MCS-SVM (selected 40 signals) with linear and RBF kernels were comparable with

| Table 1. Misclassification rates (in %) and standard errors (in italics) of different classifiers in simulated datasets. |
|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| Sparse SVM-GLM- | Sparse NN-SVM- | Sparse NN-SVM- | gSAVG WoS | gSAVG MarS | bgSAVG MixS |
| Ex | Bayes | LIN | NET | LASSO | RAND | RBF | γ1 | γ2 | γ3 | γ1 | γ2 | γ3 | γ1 | γ2 | γ3 | γ1 | γ2 | γ3 |
|----|-------|-----|-----|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1  | 15.87 | 20.20 | 17.60 | 19.70 | 44.21 | 27.01 | 21.87 | 41.37 | 36.97 | 36.24 | 20.89 | 17.45 | 17.11 | 19.94 | 17.51 | 17.10 |
| 2  | 5.07  | 49.20 | 49.98 | 49.92 | 49.45 | 50.50 | 49.37 | 44.03 | 46.94 | 46.51 | 49.96 | 49.66 | 49.54 | 49.96 | 49.66 | 49.54 |
| 3  | 4.85  | 23.80 | 23.35 | 26.65 | 46.74 | 34.30 | 31.35 | 42.65 | 41.24 | 40.57 | 22.63 | 22.56 | 22.56 | 37.32 | 21.14 | 22.28 |
| 4  | 7.50  | 47.60 | 50.11 | 49.92 | 50.19 | 49.39 | 38.59 | 40.99 | 45.17 | 45.23 | 10.14 | 18.80 | 19.09 | 12.45 | 24.23 | 33.09 |
| 5  | 11.32 | 49.40 | 46.09 | 50.31 | 49.91 | 48.06 | 47.17 | 36.55 | 34.31 | 43.24 | 12.42 | 12.07 | 50.12 | 13.42 | 12.21 | 49.99 |
| 6  | 0.14  | 0.35  | 0.71  | 0.18  | 0.18  | 0.31  | 0.31  | 0.26  | 0.23  | 0.21  | 0.15  | 0.14  | 0.21  | 0.32  | 0.14  | 0.21  |
| 7  | 13.07 | 50.60 | 49.87 | 49.87 | 49.81 | 50.23 | 48.91 | 41.90 | 33.80 | 39.95 | 17.02 | 14.35 | 49.30 | 22.43 | 14.30 | 49.28 |
| 8  | 12.83 | 50.11 | 50.17 | 49.92 | 50.04 | 50.08 | 50.16 | 48.55 | 49.36 | 49.32 | 50.13 | 49.68 | 49.54 | 36.64 | 42.25 | 45.17 |
| 9  | 0.14  | 0.21  | 0.23  | 0.22  | 0.22  | 0.25  | 0.23  | 0.22  | 0.21  | 0.21  | 0.25  | 0.22  | 0.24  | 0.27  | 0.44  | 0.40  |

NOTE: The figure in bold indicates the minimum misclassification rate.
Sparse NN-RAND, which yielded the minimum misclassification rate here. Meanwhile, Sparse SVMLIN led to a promising performance too.

With \( d = 22283 \), GSE3726 has the highest number of covariates among all the datasets, but the sample size is only 32. On an average, the number of variables selected by MarS and MixS ranged from 1.01 to 2.65 and 1.03 to 2.85, respectively. For \( \gamma_2 \) and \( \gamma_3 \), we observed considerable improvement in the misclassification rates for gSAVG with MarS when compared against gSAVG implemented without screening. As expected, MixS yielded further improvement in the performance of the bgSAVG classifier by accounting for the differences in joint distributions. In fact, bgSAVG when combined with MixS for \( \gamma_2 \) yielded the lowest misclassification rate. The MCS classifiers retained 30 components, and only they could outperform the WoS version of gSAVG. All the other classifiers failed to capture sparsity in the covariates, and led to significantly higher misclassification rates.

In the GSE967 dataset, MCS screened 7 components, while the number of components retained by MixS ranged from 4.68 to 28.7. Surprisingly, MixS with \( \gamma_1 \) screened several signal components (see Table I in Appendix C of the Supplementary). The choice \( \gamma_2 \) led to the best misclassification rate when combined with MixS and bgSAVG. Among the competitors, only the linear classifiers, namely, Sparse SVMLIN, GLMNET, and MCS-SVMLIN showed improvement when combined with the gSAVG classifier without any screening. On the other hand, the nonlinear classifiers, namely, Sparse NN-RAND and MCS-SVMRBF performed even worse than that classifier.

### 6. Discussion

Using energy distances, we have developed some methods of variable screening for classification. The novelty with the proposed method MarS is that it not only retains all the signals but also discards some of the noise variables with probability tending to 1 as \( n \to \infty \). We have also built the MixS algorithm which screens pairs of components having differences in their joint distributions. Furthermore, a discrimination criterion that is coherent with the screening method has been proposed. The classifier is shown to yield perfect classification under fairly general conditions. We have used a variety of simulated examples and some benchmarks as well as real datasets to amply demonstrate the superiority of the proposed classifiers compared to several popular classification methods.

For the MixS algorithm, computation of a \( d \times d \) matrix involving all pair-wise energy distances leads to quadratic complexity (in terms of \( d \)). This is an inherent requirement for the exploration of signals beyond the marginals, which is also in line with some recent studies (see, e.g., Song et al. 2018; Li, Hong, and Li 2019; Jiang et al. 2022). For linear models, the heredity principle (HP) (Nelder 1977) is often used to reduce complexity for screening of interaction effects (see, e.g., Bien, Taylor, and Tibshirani 2013; Hao, Feng, and Zhang 2018). In HP, an interaction effect is not included in the model if its constituents have no marginal significance. Note that HP does not hold when the variables are jointly important, but the marginal signals are weak, and the algorithms in Hao, Feng, and Zhang (2018) would break down. Our proposed PairS algorithm (without any heredity structure) addresses this problem, but in a model-free manner. For this framework, one may extend the notion of HP as follows:

- strong HP: \( F(X_k, X_j) \neq G(X_k, X_j) \Rightarrow F(X_k) \neq G(X_k) \) and \( F(X_j) \neq G(X_j) \),
- weak HP: \( F(X_k, X_j) \neq G(X_k, X_j) \Rightarrow F(X_k) \neq G(X_k) \) or \( F(X_j) \neq G(X_j) \).

Using this extended notion of HP, one may develop a two-stage method similar to the fORT algorithm proposed by Hao and Zhang (2014), leading to a significant reduction in the computational complexity of our MixS algorithm. However, developing a fast numerical algorithm, which also enjoys a comprehensive theoretical framework will require further investigation.

Another bottleneck for the MixS algorithm is where permutation tests are conducted for identifying paired features. To avoid this problem, one may think of using a parametric test instead of the permutation test. Note that the asymptotic null distribution of the energy statistic is an infinite weighted sum of independent Chi-square random variables, where the weights depend on the underlying distribution (see Székely and Rizzo 2004). This distribution has been approximated using a gamma distribution with appropriate choices of shape and scale parameters (see, e.g., Pfister et al. 2018). On a different note, a recent work by Shen, Panda, and Vogelstein (2022) approximates the
tail behavior (of an appropriately centered and scaled version) of the infinite sum by a single Chi-square random variable. However, adapting these methods to our situation is not straightforward, although it is worth investigating and is a topic of future research.

Supplementary Materials

All proofs and mathematical details are provided in the Supplementary material available online. The Supplementary also contains algorithms of the proposed screening methods and some additional numerical results.

Acknowledgments

The authors would like to thank the Editor, Associate Editor and two anonymous reviewers for their careful reading of the manuscript and thoughtful comments.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The research of the second author was partially supported by the INSPIRE Faculty Fellowship from the Department of Science and Technology, Government of India.

ORCID

Sarbojit Roy https://orcid.org/0000-0003-1659-7318
Soham Sarkar https://orcid.org/0000-0003-0412-6826
Subhajit Dutta https://orcid.org/0000-0001-8483-2012
Anil K. Ghosh https://orcid.org/0000-0002-2476-8458

References

Balakrishnan, N., and Stepanov, A. (2008), "Asymptotic Properties of the Ratio of Order Statistics," *Statistics & Probability Letters*, 78, 301–310. [451]
Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013), “Ultra-high Dimensional Feature Screening via RKHS Embeddings,” in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (Vol. 31), eds. C. M. Carvalho and P. Ravikumar, pp. 126–134, Scottsdale, Arizona, USA, PMLR. [450]
Baringhaus, L., and Franz, C. (2010), “Rigid Motion Invariant Two-Sample Tests,” *Statistica Sinica*, 20, 1333–1361. [450]
Bien, J., Taylor, J., and Tibshirani, R. (2013), "A LASSO for Hierarchical Interactions," *The Annals of Statistics*, 41, 1111–1141. [460]
Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Sefirov, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Samps, N., Dougherty, E., Wang, E., Marincova, F., Gooden, C., Lueders, J., Glattfelser, A., Pollock, P., Carpen, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000), “Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling,” *Nature*, 406, 536–540. [459]
Cheng, G., Li, X., Lai, P., Song, F., and Yu, J. (2017), "Robust Rank Screening for Ultrahigh Dimensional Discriminant Analysis," *Statistics and Computing*, 27, 535–545. [448]
Cui, H., Li, R., and Zhong, W. (2015), "Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of the American Statistical Association*, 110, 630–641. [448]
Deegealla, S., and Bostrom, H. (2006), "Reducing High-Dimensional Data by Principal Component Analysis vs. Random Projection for Nearest Neighbor Classification," in 2006 5th International Conference on Machine Learning and Applications (ICMLA06), pp. 245–250. [458]
Derigs, U. (1988), "Solving Non-bipartite Matching Problems via Shortest Path Techniques," *Annals of Operations Research*, 13, 225–261. [453]
Dutta, S., and Genton, M. G. (2014), “A Non-gaussian Multivariate Distribution with all Lower-Dimensional Gaussians and Related Families,” *Journal of Multivariate Analysis*, 132, 82–93. [454]
Fan, J., and Fan, Y. (2008), “High Dimensional Classification Using Features Annealed Independence Rules,” *The Annals of Statistics*, 36, 2605–2637. [448]
Fan, J., and Lv, J. (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space,” *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [448]
Fan, J., and Song, R. (2010), “Sure Independence Screening in Generalized Linear Models with NP-Dimensionality,” *The Annals of Statistics*, 38, 3567–3604. [449]
Feller, W. (1971), *An Introduction to Probability Theory and its Applications* (Vol. II, 2nd ed.), New York-London-Sydney: Wiley. [449]
Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005), “Measuring Statistical Dependence with Hilbert-Schmidt Norms,” in *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings* 16, pp. 63–77, Springer. [450]
Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2005). "Result Analysis of the NIPS 2003 Feature Selection Challenge," in *Advances in Neural Information Processing Systems*, eds. L. Saul, Y. Weiss, and L. Bottou, Cambridge, MA: MIT Press. [458]
Hao, N., Feng, Y., and Zhang, H. H. (2018), “Model Selection for High-Dimensional Quadratic Regression via Regularization,” *Journal of the American Statistical Association*, 113, 615–625. [449,460]
Hao, N., and Zhang, H. H. (2014), “Interaction Screening for Ultrahigh-Dimensional Data,” *Journal of the American Statistical Association*, 109, 1285–1301. [460]
Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, New York: Springer. [458]
He, S., Ma, S., and Xu, W. (2019), “A Modified Mean-Variance Feature-Screening Procedure for Ultrahigh-Dimensional Discriminant Analysis,” *Computational Statistics & Data Analysis*, 137, 155–169. [448]
Jiang, H., Zhao, X., Ma, R. C. W., and Fan, X. (2022), “Consistent Screening Procedures in High-Dimensional Binary Classification,” *Statistica Sinica*, 32, 109–130. [448,449,460]
John, G. H., Kohavi, R., and Pfleger, K. (1994), "Irrelevant Features and the Subset Selection Problem," in *Machine Learning Proceedings 1994*, pp. 121–129, Elsevier. [459]
Li, Y., Hong, H. G., and Li, Y. (2019), “Multiclass Linear Discriminant Analysis with Ultrahigh-Dimensional Features,” *Biometrics*, 75, 1086–1097. [460]
Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), “Optimal Nonbipartite Matching and its Statistical Applications,” *The American Statistician*, 65, 21–30. [455]
Mai, Q., and Zou, H. (2013), “The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification,” *Biometrika*, 100, 229–234. [448]
Meier, L., Van De Geer, S., and Bühlmann, P. (2008), “The Group Lasso for Logistic Regression,” *Journal of the Royal Statistical Society*, Series B, 70, 53–71. [449]
Nelder, J. A. (1977), "A Reformulation of Linear Models," *Journal of the Royal Statistical Society*, Series A, 140, 48–77. [460]
Ni, L., and Fang, F. (2016), “Entropy-based Model-Free Feature Screening for Ultrahigh-Dimensional Multiclass Classification,” *Journal of Non-parametric Statistics*, 28, 515–530. [455]
Pan, R., Wang, H., and Li, R. (2016), “Ultrahigh-Dimensional Multiclass Linear Discriminant Analysis by Pairwise Sure Independence Screening,” *Journal of the American Statistical Association*, 111, 169–179. [448]
Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018), “Kernel-based Tests for Joint Independence,” *Journal of the Royal Statistical Society*, Series B, 80, 5–31. [460]
Roy, S., Sarkar, S., Dutta, S., and Ghosh, A. K. (2022), "On Generalizations of Some Distance based Classifiers for HDLSS Data," *Journal of Machine Learning Research*, 23, 1–41. [457]

Shen, C., Panda, S., and Vogelstein, J. T. (2022), "The Chi-square Test of Distance Correlation," *Journal of Computational and Graphical Statistics*, 31, 254–262. [460]

Sheng, Y., and Wang, Q. (2020), "Model-Free Feature Screening for Ultrahigh Dimensional Classification," *Journal of Multivariate Analysis*, 178, 104618. [448]

Ship, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., et al. (2002), "Diffuse Large b-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning," *Nature Medicine*, 8, 68–74. [459]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [449,458]

Song, F., Lai, P., Shen, B., and Cheng, G. (2018), "Variance Ratio Screening for Ultrahigh Dimensional Discriminant Analysis," *Communications in Statistics. Theory and Methods*, 47, 6034–6051. [448,449,460]

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012), "Feature Selection via Dependence Maximization," *Journal of Machine Learning Research*, 13, 1393–1434. [450]

Szekely, G. J., and Rizzo, M. L. (2004), "Testing for Equal Distributions in High Dimension," *InterStat*, 5, 1249–1272. [454,460]

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [450]

Wang, Z., Deng, G., and Xu, H. (2023), "Group Feature Screening based on Gini Impurity for Ultrahigh-Dimensional Multi-Classification," *AIMS Mathematics*, 8, 4342–4362. [449]

Yi, C., and Huang, J. (2017), "Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression," *Journal of Computational and Graphical Statistics*, 26, 547–557. [458]

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011), "Model-Free Feature Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1475. [448]