Statistica Sinica **32** (2022), 109-130
doi:https://doi.org/10.5705/ss.202020.0088

CONSISTENT SCREENING PROCEDURES IN HIGH-DIMENSIONAL BINARY CLASSIFICATION

Hangjin Jiang[1,2], Xingqiu Zhao[3], Ronald C.W. Ma[2], Xiaodan Fan[2]

[1] *Zhejiang University,* [2] *The Chinese University of Hong Kong and* [3] *The Hong Kong Polytechnic University*

Abstract: We consider variable screening in high-dimensional binary classification. First, we propose nonparametric test statistics for the problem of the two-sample distribution comparison. These test statistics combine the merits of the chi-squared and Kolmogorov–Smirnov statistics, and provide new insights into the equality test of the unspecified distributions underlying the two independent samples. Based on our new statistics, we propose a marginal screening procedure and a pairwise joint screening procedure for detecting important variables in high-dimensional binary classification. Both screening procedures have the consistent screening property, which is stronger than the sure screening property of most existing methods. The marginal screening procedure is much more powerful than other methods over a broad range of cases, and the pairwise joint screening procedure provides a way of detecting variables with a joint effect, but no marginal effect. Extensive simulations and a real-data application show the effectiveness and advantages of the proposed methods.

Key words and phrases: Binary classification, consistency, non-parametric test, Two-sample distribution comparison, variable screening.

1. Introduction

Variable screening aims to screen important variables out of thousands of candidates, and is a fundamental statistical problem in many applied areas. For example, in case-control disease studies, researchers want to find important disease factors out of numerous environmental, clinical, epigenetic, or gene expression variables. For continuous responses, many variable screening methods have been proposed; see, for example, Fan and Li (2001), Fan and Lv (2008), Fan, Feng and Song (2012), Hall and Miller (2012), Huang and Zhu (2016), Li, Zhong and Zhu (2012b), Li et al. (2012a), and the references therein. Fewer methods have been proposed for the binary response case, and include the marginal t-test screening (Fan and Fan (2008)), maximum marginal likelihood screening

Corresponding author: Xiaodan Fan, Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. E-mail: xfan@cuhk.edu.hk.

(Fan and Song (2010)), and Kolmogorov Filter (KF), based on the Kolmogorov–Smirnov (K–S) statistic (Mai and Zou (2013)). The method proposed by Cui, Li and Zhong (2017) can be seen as a generalization of the KF to the multi-class problem. These methods all screen variables according to their marginal effect, which means they may lose important variables that are marginally undetectable, but jointly detectable. Furthermore, the K–S test has low statistical power in detecting densities with bumps or high-frequency components (Fan (1996); Eubank and LaRiccia (1992)). As a result, the KF may lose some important variables, even if they are marginally detectable. Finally, the model size produced by the KF is difficult to interpret in terms of true positives and false positives. These issues motivate us to propose new screening methods for binary classification that are consistent and have a screening threshold that directly controls the false positive rate. Essentially, we reconsider the following two-sample distribution testing problem, based on independent observations of two continuous R^d-valued random vectors, $\mathbf{X}$ and $\mathbf{Y}$:

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G, \tag{1.1}$$

where F and G are the distribution functions of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $d \geq 1$ is a positive integer. This fundamental statistical testing problem has received considerable attention in the statistical literature, and has a wide range of applications (see Thas (2010) for a review).

For the univariate case ($d = 1$), classical tests such as the K–S test, Cramér-von Mises (CvM) criterion, and Anderson–Darling (A–D) statistic (Darling (1957)) are widely used. However, they usually suffer from low power when detecting densities containing high-frequency components or local features such as bumps (Fan (1996); Eubank and LaRiccia (1992)). To deal with these problems, smoothing-based tests (Neyman (1937); Fan (1996); Bera, Ghosh and Xiao (2013)) have been shown to be more powerful than classical tests over a broad range of realistic alternatives.

The multivariate case ($d > 1$) has also been studied extensively; see, for example, Weiss (1960), Friedman and Rafsky (1979), Schilling (1986), Henze (1988), Hall and Tajvidi (2002), Ludwig and Carsten (2004), Rosenbaum (2005), Ludwig and Carsten (2010), Székely and Rizzo (2013), Biswas and Ghosh (2014), Chen and Friedman (2017), Kim, Balakrishnan and Wasserman (2020), and the references therein. Most existing tests are nonparametric. For a review and numerical comparison of these tests, refer to Biswas and Ghosh (2014), who find that most tests have poor power in high-dimensional settings, but that the method of

Ludwig and Carsten (2004) performs well. Kim, Balakrishnan and Wasserman (2020) generalized the CvM statistic (denoted as gCvM) to the multivariate case by using projection-averaging, and established asymptotic theories for the proposed statistics based on U-statistic theory. The gCvM has good scalability when the dimension increases, but it inherits the weakness of the CvM. Thus it has low power when detecting densities containing high-frequency components or local features such as bumps.

Zhou et al. (2017) recently developed a new two-sample smoothing-based test that outperforms the K–S test, CvM criterion, and smooth test of Bera, Ghosh and Xiao (2013) in univariate settings, and outperforms the method of Ludwig and Carsten (2004) in multivariate settings. However, their method has three main limitations: (a) their complex parametric assumption is relatively restrictive and not easy to check in real applications; (b) the number of orthogonal directions used to construct the test statistic is critical, but difficult to determine; and (c) the optimization problem is not easy to solve for high-dimensional cases.

We first propose a new class of nonparametric test statistics, namely, the maximum adjusted chi-squared (MAC) statistics, for the two-sample distribution comparison problem (1.1). The proposed tests have the following advantages: (a) they are consistent for all kinds of continuous alternatives (see Theorem 1); (b) their finite-sample performance is better than that of existing methods in many cases, as shown in simulation studies; and (c) they are straightforward to compute, without complicated optimization.

Based on the MAC statistics, we propose consistent and model-free screening procedures for variable screening in ultrahigh-dimensional binary classification. The new marginal screening procedure has three major advantages over existing methods: (1) it enjoys the consistent screening property, instead of the sure screening property; (2) the screening threshold can be chosen to explicitly control false positives; (3) it is more powerful than other methods, as shown in the simulations in Section 4.2. In addition, our new pairwise joint screening procedure enables us to find variables that are jointly detectable, but marginally undetectable.

The rest of the paper is organized as follows. In Section 2, we introduce the new MAC statistics for the two-sample distribution comparison, and establish their consistency. In Section 3, based on the MAC statistics, we propose new procedures for variable screening in ultrahigh-dimensional binary classification, and establish their consistent screening property. In Section 4, we present our simulation studies and a real-data application. Section 5 concludes the paper. Additional simulations and all proofs are provided in the online **Supplementary**

Material.

2. Two-Sample Distribution Comparison

In this section, we present the proposed MAC test statistics for the two-sample distribution comparison under different settings. Before that, we introduce some notation. Let $\mathbf{x} = \{\mathbf{x}_i \in R^d, i = 1, 2, \ldots, n\}$ be n independent and identically distributed (i.i.d.) observations of $\mathbf{X} \in R^d$ following an unknown distribution F, and let $\mathbf{y} = \{\mathbf{y}_i \in R^d, i = 1, 2, \ldots, m\}$ be m i.i.d. observations of $\mathbf{Y} \in R^d$ following the unknown distribution G. Here, n may not be equal to m. Our goal is to test the problem given in (1.1) based on these two independent samples, $\mathbf{x}$ and $\mathbf{y}$. Let $d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}$ be the Euclidean distance between any two vectors $\mathbf{a} \in R^d$ and $\mathbf{b} \in R^d$, and let $I(\cdot)$ be the indicator function.

2.1. Univariate case: $d = 1$

The MAC statistics are motivated by two key observations. The first comes from the chi-squared statistic for the goodness-of-fit test in the one-sample case, that is, testing $F = G$ with a known continuous distribution G. The chi-squared statistic is defined as $X^2 = \sum_{i=1}^{k}(np_i - nq_i)^2/nq_i$, where p_i is the estimated probability of the event $\{X \in A_i\}$, q_i is the true probability of this event under H_0, and $\{A_1, \ldots, A_k\}$ is a partition of the support of G. Although this statistic can be generalized to the two-sample test problem (1.1), it is difficult to find the optimal k and optimal partition in real applications. A larger or smaller k, relative to the optimal value, may lead to a test with lower power. Our MAC statistic avoids these problems by focusing on local bi-partitions around sample points instead of globally dividing the support into k partitions.

The second motivating observation comes from the K–S test statistic, KS(F, G) $= \sup_{x \in R} |F(x) - G(x)|$. Computing the K–S statistic requires two steps: (1) compute the cumulative difference at each sample point x between two distributions according to the partition $(-\infty, x)$; and (2) determine the maximal difference over all sample points. However, focusing only on a single partition $(-\infty, x)$ leads to the low-resolution problem, which results in the weakness in detecting local features such as bumps and high-frequency components (Fan (1996)). To address this problem, our MAC statistic scans all data-dependent partitions at each sample point, which increases the resolution.

Given any observations x_0 and y_0 of $\mathbf{X}$ and $\mathbf{Y}$, respectively, we define $A_1 = A_{x_0,y_0} = \{x \in R : d(x, x_0) \leq d(x_0, y_0)\}$ and $A_2 = A_1^c$, that is, the complement of A_1. Define $P_i = \sum_{j=1}^{n} I(x_j \in A_i)$, $Q_i = \sum_{j=1}^{m} I(y_j \in A_i)$, and $R_i = P_i + Q_i$, for

$i = 1, 2$. Then, we construct the local statistic at (x_0, y_0) as

$$T_1(x_0, y_0) = \sum_{i=1}^{2} \frac{(P_i - (n/(n+m))R_i)^2}{(n/(n+m))R_i} + \frac{(Q_i - (m/(n+m))R_i)^2}{(m/(n+m))R_i}.$$

In the above formula, if $R_i = 0$ in a denominator, we define the corresponding ratio term as zero. Lemma S.1 in Section S4 in the **Supplementary Material** shows that $T_1(x_0, y_0) \to \chi_1^2$, as $n \to \infty$, when the null hypothesis in the problem (1.1) is true. Thus, it is called an adjusted chi-squared statistic. According to the definition of A_1 and A_2, $T_1(x_0, y_0)$ measures the distribution difference in a neighborhood of x_0. By changing the value of y_0, $T_1(x_0, y_0)$ scans over different neighborhoods of x_0, checking the distribution difference under different resolutions. On the other hand, by changing the value of x_0, $T_1(x_0, y_0)$ scans the distribution difference at different locations. The MAC statistic for the problem (1.1) is defined as the maximum of all local statistics:

$$\mathrm{MAC}_1(X, Y) = \max_{1 \le i \le n, 1 \le j \le m} \max\{T_1(x_i, y_j), T_1(y_j, x_i)\}. \tag{2.1}$$

We reject the null hypothesis in the problem (1.1) when $\mathrm{MAC}_1(X, Y) > c_0$, where c_0 is a positive threshold. In the notation MAC_1, the subscript 1 indicates that it applies to the one-dimensional case. Its consistency is studied in Section 2.4.

2.2. Two-dimensional case: $d = 2$

For the two-dimensional case, similarly to the univariate case, we first define the corresponding local test statistic at each sample point, and then take their maximum.

Given any observations $\mathbf{x}_0 = (x_{10}, x_{20})$ and $\mathbf{y}_0 = (y_{10}, y_{20})$ of $\mathbf{X}$ and $\mathbf{Y}$, respectively, we define $A_{\mathbf{x}_0,\mathbf{y}_0} = \{\mathbf{x} = (x_1, x_2) \in R^2 : d(x_1, x_{10}) \le d(x_{10}, y_{10})\}$ and $B_{\mathbf{x}_0,\mathbf{y}_0} = \{\mathbf{x} = (x_1, x_2) \in R^2 : d(x_2, x_{20}) \le d(x_{20}, y_{20})\}$. Similarly to the one-dimensional case, we check the distribution difference at different locations by changing the value of $\mathbf{x}_0$, and check the difference at different resolutions by changing the value of $\mathbf{y}_0$. Define $A_{11} = A_{\mathbf{x}_0,\mathbf{y}_0} \bigcap B_{\mathbf{x}_0,\mathbf{y}_0}$, $A_{12} = A^c_{\mathbf{x}_0,\mathbf{y}_0} \bigcap B_{\mathbf{x}_0,\mathbf{y}_0}$, $A_{21} = A_{\mathbf{x}_0,\mathbf{y}_0} \bigcap B^c_{\mathbf{x}_0,\mathbf{y}_0}$, $A_{22} = A^c_{\mathbf{x}_0,\mathbf{y}_0} \bigcap B^c_{\mathbf{x}_0,\mathbf{y}_0}$, $P_{ij} = \sum_{k=1}^{n} I(\mathbf{x}_k \in A_{ij})$, $Q_{ij} = \sum_{k=1}^{m} I(\mathbf{y}_k \in A_{ij})$, and $R_{ij} = P_{ij} + Q_{ij}$, for $i = 1, 2$ and $j = 1, 2$, the local adjusted chi-squared statistic at $(\mathbf{x}_0, \mathbf{y}_0)$ is given by

$$T_2(\mathbf{x}_0, \mathbf{y}_0) = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(P_{ij} - (n/(m+n))R_{ij})^2}{(n/(m+n))R_{ij}} + \frac{(Q_{ij} - (m/(m+n))R_{ij})^2}{(m/(m+n))R_{ij}}.$$

If any $R_{ij} = 0$, we define the corresponding ratio term in the above formula as zero. We show in Lemma S.1 in Section S4 in the **Supplementary Material** that $T_2(x_0, y_0) \to \chi^2_3$, as $n \to \infty$, when the null hypothesis in problem (1.1) is true.

The MAC test statistic for the two-dimensional case in problem (1.1) is defined as

$$\mathrm{MAC}_2(\mathbf{X}, \mathbf{Y}) = \max_{1\le i\le n, 1\le j\le m} \max\{T_2(\mathbf{x}_i, \mathbf{y}_j), T_2(\mathbf{y}_j, \mathbf{x}_i)\}. \tag{2.2}$$

We reject H_0 in the problem (1.1) if $\mathrm{MAC}_2(\mathbf{X}, \mathbf{Y}) > c_0$, where c_0 is a positive threshold. Its consistency is studied in Section 2.4.

2.3. Multi-dimensional case: $d > 2$

The way to construct the test statistic for the multi-dimensional case is a bit different from the two-dimensional case. We first transform the problem into a two-dimensional problem by partitioning the d variables into two groups, and then apply the idea for the two-dimensional case. Let $S = \{1, 2, \ldots, d\}$ be the index set. For any nonempty set $s \subsetneqq S$, we define $\mathbf{X}_s$ as the corresponding subset of $\mathbf{X}$. For example, when $s = \{1, 2\}$, $\mathbf{X}$ is partitioned into two groups, as $\mathbf{X}_s = (X_1, X_2)$ and $\mathbf{X}^c_s = (X_3, X_4, \ldots, X_d)$.

Now, we construct the test statistic for the multi-dimensional case. Given any nonempty set $s \subsetneqq S$, any observation $\mathbf{x}_0 = (x_{01}, \ldots, x_{0d})$ of $\mathbf{X}$ and $\mathbf{y}_0 = (y_{01}, \ldots, y_{0d})$ of $\mathbf{Y}$, we first get the grouping $\mathbf{x}_0 = (\mathbf{x}_{0s}, \mathbf{x}^c_{0s})$ and $\mathbf{y}_0 = (\mathbf{y}_{0s}, \mathbf{y}^c_{0s})$, and then define the local test statistic at $(\mathbf{x}_0, \mathbf{y}_0)$ as

$$T_s(\mathbf{x}_0, \mathbf{y}_0) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(P^s_{ij} - (n/(m+n))R^s_{ij})^2}{(n/(m+n))R^s_{ij}} + \frac{(Q^s_{ij} - (m/(m+n))R^s_{ij})^2}{(m/(m+n))R^s_{ij}},$$

with P^s_{ij}, Q^s_{ij} and R^s_{ij} defined according to $(\mathbf{x}_0, \mathbf{y}_0)$ in a way similar to P_{ij}, Q_{ij}, and R_{ij} in the two-dimensional case. If $R^s_{ij} = 0$, we define the corresponding ratio term as zero.

The MAC test statistic for the multi-dimensional case in problem (1.1) is defined as

$$\mathrm{MAC}_3(\mathbf{X}, \mathbf{Y}) = \max_{s\neq\emptyset, s\subsetneqq S} \max_{1\le i\le n, 1\le j\le m} \max\{T_s(\mathbf{x}_i, \mathbf{y}_j), T_s(\mathbf{y}_j, \mathbf{x}_i)\}. \tag{2.3}$$

We reject H_0 in (1.1) if $\mathrm{MAC}_3(\mathbf{X}, \mathbf{Y}) > c_0$, where c_0 is a positive threshold. The consistency of this test is studied in Section 2.4.

Remark 1. Note that the asymptotic distribution of $T_s(\mathbf{x}_0, \mathbf{y}_0)$ under H_0 does

not depend on s when the sample sizes go to infinity (see Lemma S.1 in Section S4 in the **Supplementary Material**). This is key to the consistency of MAC_3.

Remark 2. Although we have proposed MAC_3 for problem (1.1) in the high-dimensional case, and its performance is much better than other methods (see the simulations in the **Supplementary Material**), it is, in general, not a good choice for $d > 10$ because of its computational inefficiency. Thus, it is still interesting to investigate how to define more computationally efficient and powerful test statistics for problem (1.1) in the high-dimensional case.

2.4. Consistency

To establish the consistency of the proposed test statistics, we need the following assumption, which requires that the sample sizes of these two samples be comparable.

Assumption 1. *There exists a positive number C, such that $1/C \le n/m \le C$.*

This assumption excludes extremely unbalanced cases, and ensures that the sample sizes n and m go to infinity at the same rate. The consistency of MAC_i, for $i = 1, 2, 3$, is established in the following theorem, which states that as long as there are sufficient observations, we always reject the null hypothesis if a fixed alternative hypothesis is true. The proof of the theorem is given in Section S4 in the **Supplementary Material**.

Theorem 1. *Assume that $\{\boldsymbol{x}_i, i = 1, 2, \ldots, n\}$ and $\{\boldsymbol{y}_i, i = 1, 2, \ldots, m\}$ are independent observations of the continuous random variables $\boldsymbol{X} \in R^d$ and $\boldsymbol{Y} \in R^d$, respectively. If there exists a positive number C such that $1/C \le n/m \le C$, then for MAC_i $(i = 1, 2, 3)$, defined in Equations (2.1)–(2.3), we have the following:*

(I). *Under H_0, as $n, m \to +\infty$, the following inequalities hold with probability going to one: (a) $\mathrm{MAC}_1(X, Y) < 8\log(2nm) + 1$; (b) $\mathrm{MAC}_2(\boldsymbol{X}, \boldsymbol{Y}) < 8\log(2nm) + 3$; (c) $\mathrm{MAC}_3(\boldsymbol{X}, \boldsymbol{Y}) < 8\log(2^{d+1}nm) + 3$.*

(II). *Under H_1, there exists a positive constant c_i, such that $\mathrm{MAC}_i(\boldsymbol{X}, \boldsymbol{Y}) > c_i(n + m)$, for $i = 1, 2, 3$.*

Theorem 1 tells us that MAC_i, for $i = 1, 2, 3$, has an upper bound of order $O(\log(nm))$ under H_0, which is much smaller than its lower bound of order $O(n+m)$ under H_1. Thus, all of these tests are consistent. When the dimensions of $\mathbf{X}$ and $\mathbf{Y}$ grow with the sample size $N = n+m$, say, $d = d_N = O(N^\alpha)$, MAC_3 is still consistent if $\alpha < 1$. That is, it is consistent for the cases where the dimension d_N grows more slowly than N, that is, $d_N/N \to 0$. The computation of MAC_3 will

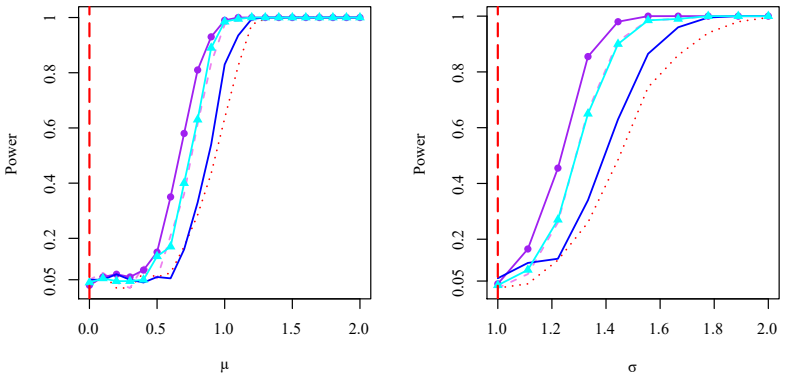

Figure 1. Power of K–S (·····), A–D (- - - -), CvM (——), MAC_1 (—●—), and **ZZZ** (——) for the Gaussian example based on 2,000 replications with significance level 0.05. The vertical dashed line (- - - -) corresponds to the case when the null hypothesis is true.

be time consuming when d is large, say, $d > 10$. However, for relatively lower-dimensional problems, MAC_3 can be computed in a straightforward manner.

To show the intuitions behind the limiting distribution of the proposed statistics, we show the empirical distribution of MAC_1 and MAC_2 under the null hypothesis when $n = m = 200$ in Figure S4 in **Supplementary Material**. It is difficult to obtain analytically the limit distributions of the MAC statistics under the null hypothesis, owing to the unknown dependence structure among the local statistics. Thus, the corresponding p-value is computed using a K-times permutation. In this study, we set K as 1,000.

The following Gaussian example illustrates the performance of MAC_1, comparing it with that of the K–S test, A–D statistic and CvM, as well as the latest method proposed in Zhou et al. (2017) (denoted as **ZZZ**). This example fits for the K–S test, A–D statistic and CvM because there are neither bumps nor high-frequency components. The performance of the MAC statistics on examples with bumps and high-frequency components is shown in Section S1 in the **Supplementary Material**, which also shows the advantage of MAC_3 over **ZZZ** in multi-dimensional cases.

Gaussian Example. (1) $X \sim F = 0.5N(-\mu, 1) + 0.5N(\mu, 1)$ and $Y \sim G = N(0, 1)$, with μ ranging from zero to two. (2) $X \sim F = N(0, \sigma^2)$ and $Y \sim G = N(0, 1)$, with σ ranging from one to two.

In this example, we set $n = m = 200$, and the power of each method is estimated based on 2,000 independent replications with significance level 0.05. The results under various parameters are shown in Figure 1. All five methods

control the type-I error rate at the targeted significance level of 0.05. MAC_1 performed the best in these two examples. The performance of A–D is similar to that of **ZZZ** in these two examples, and the performance of the K-S test is the worst. These results show the lower power of the K–S test, and thus the weakness of the KF for variable screening in binary classification. They also suggest the potential power of MAC_1 in the later problem.

3. Variable Screening

In this section, we consider variable screening in binary classification. First, we introduce some additional notation. For any set A, $|A|$ denotes the number of elements it contains. Let $Y \in \{0, 1\}$ be a binary response, n be the sample size of $Y = 0$, and m be the sample size of $Y = 1$. Let $S_1 = \{Z_j : 1 \le j \le p\}$ be the set of input variables, and $S_1^* = \{Z_j : P(Z_j^0) \neq P(Z_j^1)\}$ be the set of variables with a marginal effect, where $q = |S_1^*| \ll p$. For any random variable U, denote the two conditional random variables of U, $(U|Y = 0)$ and $(U|Y = 1)$, as U^0 and U^1, respectively.

3.1. Marginal screening

It is well known that covariate Z is associated with the binary response Y if Z^0 and Z^1 have different distributions. The KF uses the K–S statistic to measure the distance between these two distributions, and claims that the covariate with a larger distance is an important variable. However, the K–S statistic is powerless in many cases, as discussed previously and shown in the simulations in Section S1 of the **Supplementary Material**, which means the KF may miss important variables. In addition, the model size from the KF is difficult to determine. Usually, it provides a set of potentially important variables, which should be further refined in the formal model building step. There are two main concerns in this screening and refining framework: (1) if the KF provides a model size smaller than the true model size, one may lose true variables in this screening step; and (2) the refining step relies on strong model assumptions, which may lead to the loss of important variables, owing to the model misspecification. In the following, we propose a new marginal screening procedure to overcome these drawbacks. As suggested by Theorem 1, we define the new marginal screening procedure as

$$M_1(c) = \{Z_j : \mathrm{MAC}_1(Z_j^1, Z_j^0) \ge c\}, \tag{3.1}$$

where c is a threshold. Because it is based on the MAC test statistic MAC_1, we call it the MAC_1 filter (MAC_1-F). In the following, we establish the consistent

screening property of MAC_1-F, which is a more desirable property than the sure screening property of existing screening procedures. Its proof is given in Section S4 of the **Supplementary Material**.

Corollary 1. *Following the notation defined above, let* $S_1^{*c} = S_1 \backslash S_1^*$. *When assumption A holds,* $\max_{Z_j \in S_1^{*c}} \text{MAC}_1(Z_j^1, Z_j^0) < \text{MAC}_1^1(n,m) \equiv 8\log(2pnm) + 1$ *holds with probability going to one, as* $n, m \to +\infty$. *Furthermore, if* $p = e^{(n+m)^\eta}$ *with* $0 < \eta < 1$, *for the screening procedure in Equation* (3.1), *we have* $P(M_1(\text{MAC}_1^1(n,m)) = S_1^*) \to 1$ *as* $n, m \to +\infty$.

Remark 3. Denote w as the minimal positive value such that $\text{MAC}_1(Z_j^1, Z_j^0) > (n+m)w$, for all $Z_j \in S_1^*$. As long as $wn > 8\log(2pnm) + 1$, the consistent screening property of MAC_1-F ($M_1(c)$), according to Theorem 1, actually holds for any $c \in [8\log(2pnm) + 1, w(n+m))$.

As one can see from the proof of Corollary 1 (see Section S4 in the **Supplementary Material**) and Figure S1, the threshold $\text{MAC}_1^1(n,m)$ is usually too large for finite sample problems. Thus we suggest setting the threshold in the screening procedure as the $(1-\alpha)$ quantile, $\text{MAC}_1^\alpha(n,m)$, of MAC_1 under H_0, with sample sizes n and m. This procedure is denoted as

$$M_1^\alpha = \{Z_j : \text{MAC}_1(Z_j^1, Z_j^0) \geq \text{MAC}_1^\alpha(n,m)\}. \tag{3.2}$$

Thus, our procedure can determine the threshold according to a required false positive rate. This procedure indeed control the false positive rate at α when all variables are independent. In contrast, the KF and other existing screening methods provides only the order of the threshold or an ad hoc default model size. Thus they may produce unexpected false positives.

3.2. Pairwise joint screening

Both MAC_1-F and the KF screen variables according to their marginal effect, and thus may miss associated variables that are marginally undetectable. Here, we consider pairwise joint-effect screening. Similarly to marginal-effect screening, it is natural to claim the joint effect of two variables Z_i and Z_j on a binary response Y if they satisfy the inequality $P((Z_i, Z_j)|Y=0) \neq P((Z_i, Z_j)|Y=1)$; that is, they have different conditional joint distributions. However, a marginally associated variable and a non-associated variable may also lead to this inequality, thus producing false discoveries. There are four conditions of Z_i and Z_j that lead to this inequality: (J1) one variable is marginally detectable, and the other is associated with Y, but marginally undetectable; (J2) one variable is

marginally detectable, and the other is not associated with Y; (J3) both variables are marginally undetectable, but are jointly detectable; and (J4) both variables are marginally detectable. Marginal screening methods, such as MAC_1-F, can read out variables in (J4), but miss both the variables in (J3) and the marginally undetectable variable in (J1). On the other hand, selecting variable pairs solely according to $P((Z_i, Z_j)|Y = 0) \neq P((Z_i, Z_j)|Y = 1)$ mistakenly selects the non-associated variable in (J2). Thus, a desired pairwise joint screening procedure should include pairs in (J1) and (J3), and should exclude the non-associated variable in (J2).

To retrieve the variable pairs satisfying condition (J3), we propose a new screening procedure, as follows:

$$M_{21}(c) = \{X_{ij} = (Z_i, Z_j) : Z_i, Z_j \in M_1^c \text{ and } \text{MAC}_2(X_{ij}^1, X_{ij}^0) > c\}, \tag{3.3}$$

where $M_1^c = S_1 \backslash M_1$, with M_1 given by Equation (3.1), and c is a screening threshold. This procedure is called the MAC_2 Filter 1 (MAC_2-F_1).

Now, we consider the method for retrieving the variable pairs satisfying condition (J1). Assume that Z_i is marginally detectable and Z_j is not. We test $P(F^1(Z_i), Z_j|Y = 1) = P(F^0(Z_i), Z_j|Y = 0)$, where F^1 and F^0 are the cumulative distribution functions of Z_i^1 and Z_i^0, respectively. Because $U_{Z_i}^1 = (F^1(Z_i)|Y = 1)$ and $U_{Z_i}^0 = (F^0(Z_i)|Y = 0)$ are uniformly distributed random variables, Z_j is associated with Y if $P((U_{Z_i}^1, Z_j)|Y = 1) \neq P(U_{Z_i}^0, Z_j|Y = 0)$. Therefore, we define the screening procedure for this case as follows:

$$\begin{aligned} M_{22}(c) = \{X_{ij} = (Z_i, Z_j) : Z_i \in M_1, Z_j \in M_1^c \\ \text{and } \text{MAC}_2(X_{ij,F}^1, X_{ij,F}^0) > c\}, \end{aligned} \tag{3.4}$$

where $X_{ij,F}^1 = ((\hat{F}_1(Z_i), Z_j)|Y = 1)$, $X_{ij,F}^0 = ((\hat{F}_0(Z_i), Z_j)|Y = 0)$, $\hat{F}_1$ and $\hat{F}_0$ are the empirical cumulative distribution functions of Z_i^1 and Z_i^0, respectively, and c is a screening threshold. This procedure is called the MAC_2 Filter 2 (MAC_2-F_2). Note that M_{22} excludes automatically the non-associated variable in the condition (J2).

To establish the consistent screening property of M_{21} and M_{22}, we define

- $S_{21}^* = \{(Z_i, Z_j) : Z_i, Z_j \in S_1^{*c} \text{ and } P(Z_i, Z_j|Y = 0) \neq P(Z_i, Z_j|Y = 1), 1 \leq i < j \leq p\}$;

- $S_{22}^{*}=\{(Z_i,Z_j): Z_i\in S_1^{*c}, Z_j\in S_1^{*} \text{ and } P(Z_i,Z_j|Y=0)\neq P(Z_i,Z_j|Y=1)\}$;
- $S_{21}=\{(Z_i,Z_j): Z_i,Z_j\in S_1^{*c}, 1\le i<j\le p\}$;
- $S_{22}=\{(Z_i,Z_j): Z_i\in S_1^{*c}, Z_j\in S_1^{*}\}$;
- $S_{21}^{*c}=S_{21}\backslash S_{21}^{*}$;
- $S_{22}^{*c}=S_{22}\backslash S_{22}^{*}$.

Based on this notation, we have the following results; see Section S4 in the **Supplementary Material** for the proof.

Corollary 2. *Following the notation given above, when assumption A holds, we have that (a)* $\max_{X_{ij}=(Z_i,Z_j)\in S_{21}^{*c}}\mathrm{MAC}_2(X_{ij}^1,X_{ij}^0)<\mathrm{MAC}_2^1(n,m)=8\log(2p^2nm)+3$ *and (b)* $\max_{X_{ij}=(Z_i,Z_j)\in S_{22}^{*c}}\mathrm{MAC}_2(X_{ij,F}^1,X_{ij,F}^0)<\mathrm{MAC}_2^2(n,m)=8\log(2pqnm)+3$ *hold with probability going to one as* $n,m\to+\infty$. *Furthermore, if* $p=e^{(n+m)^\eta}$, *with* $0<\eta<1$, *for the screening procedure in Equation* (3.3) *and* (3.4), *we have* $P(M_{21}(\mathrm{MAC}_2^1(n,m))=S_{21}^{*})\to 1$ *and* $P(M_{22}(\mathrm{MAC}_2^2(n,m))=S_{22}^{*})\to 1$ *as* $n,m\to+\infty$.

Remark 4. The consistency of $M_{21}(c)$ and $M_{22}(c)$ can be understood in a similar way to that of $M_1(c)$ (see *Remark* 3).

Similarly to MAC_1-F, in practice, we set the thresholds for M_{21} and M_{22} as the $(1-\alpha)$ quantile, $\mathrm{MAC}_2^{\alpha}(n,m)$, of MAC_2 under H_0. The corresponding procedures are specified by

$$\begin{aligned} M_{21}^{\alpha}=\{&X_{ij}=(Z_i,Z_j): Z_i,Z_j\in M_1^{c}\\ &\text{and } \mathrm{MAC}_2(X_{ij}^1,X_{ij}^0)>\mathrm{MAC}_2^{\alpha}(n,m)\}, \qquad (3.5)\\ M_{22}^{\alpha}=\{&X_{ij}=(Z_i,Z_j): Z_i\in M_1, Z_j\in M_1^{c}\\ &\text{and } \mathrm{MAC}_2(X_{ij,F}^1,X_{ij,F}^0)>\mathrm{MAC}_2^{\alpha}(n,m)\}. \qquad (3.6)\end{aligned}$$

In summary, we propose a three-step procedure for pairwise joint screening: (1) use MAC_1-F to select M_1; (2) use MAC_2-F_1 to select variables pairs satisfying condition (J3); and (3) use MAC_2-F_2 to select the marginal nondetectable variable satisfying condition (J1). We call this three-step procedure the MAC filter (MAC-F).

4. Numerical Studies

In this section, we explore the power of our new screening procedures through simulations. In the first part, we compare MAC_1-F with KF because both are marginal screening methods. In the second part, we show the power of our pairwise joint screening method, MAC-F, by comparing it with the marginal screening procedure MAC_1-F.

4.1. Marginal screening

In this section, we compare the performance of our new screening procedure MAC_1-F and that of the KF using five examples. Examples 1–4 are designed to show the weakness of the KF in cases where the K–S test is powerless. Example 5 is taken from Mai and Zou (2013) to show the comparable performance of MAC_1-F when the KF performs well. Other examples from Mai and Zou (2013) are also tested (see Section S2 in the **Supplementary Material**). In these simulations, we follow Mai and Zou (2013) by setting $p = 2,000$ and $n = m = 200$. For each case, 500 independent experiments are performed. Because the KF has difficulty in choosing the screening threshold, the smallest model size required to contain all the true variables is used to evaluate the performance of these two methods (Mai and Zou (2013); Li, Zhong and Zhu (2012b)).

Example 1.

- $X_j|Y = 1 \sim \text{uniform}(-1, 1)$, $X_j|Y = 0 \sim g_c(x) = 0.5 + 0.5\sin(2\pi cx)$, with $c = 1.5$ and $j = 1, \ldots, 5$.
- $X_j : j = 6, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$.

Example 2.

- $X_j|Y = 1 \sim f(x) = \text{lognormal}(0, 1)$, $X_j|Y = 0 \sim g_c(x) = f(x)(1 + c\sin(2\pi \log x))$, with $c = 1$ and $j = 1, \ldots, 5$.
- $X_j : j = 6, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$.

Example 3.

- $X_j|Y = 1 \sim \text{uniform}(0, 1)$, $X_j|Y = 0 \sim g_c(x) = \exp\{c\sin(5\pi x)\}$, with $c = 1.5$ and $j = 1, \ldots, 5$.
- $X_j : j = 6, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$.

Example 4.

- $X_j|Y = 1 \sim \text{uniform}(0, 1)$, $X_j|Y = 0 \sim g_c(x) = 1 + c\cos(5\pi x)$, with $c = 1.5$ and $j = 1, \ldots, 5$.

Table 1. Smallest model size required to contain all true variables for sample sizes $n = m = 200$. The numbers are medians from 500 replicates with the standard errors (estimated by bootstrap) given in parentheses.

Method	Example 1	Example 2	Example 3	Example 4	Example 5
KF	36 (2.2)	679(28.7)	210(9.9)	280(6.0)	5(0)
MAC_1-F	14.0(0.6)	26(1.3)	5(0)	5(0)	5(0)

- $X_j : j = 6, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$

Example 5.

- $X_j|Y = 1 \sim t_4$, $X_j|Y = 0 \sim 0.5N(2.5, 1) + 0.5N(-2.5, 1)$, for $j = 1, \ldots, 5$.
- $X_j : j = 6, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$.

The simulation results are shown in Table 1. As expected, for Examples 1–4, where the K–S test has relatively lower power as showed in Examples 2–5 in the **Supplementary Material**, the KF selects more noisy variables than MAC_1-F does. For Example 5, the two methods perform similarly. The results given in Table S4 in the **Supplementary Material** show that MAC_1-F and the KF both perform well for other examples from Mai and Zou (2013). More details are provided in Section S2.2 in the **Supplementary Material**. In addition, we show the simulation results for smaller sample sizes in Tables S1–S3 of Section S2.1 of the **Supplementary Material**, which show that MAC_1-F outperforms the KF when the sample sizes are smaller.

A key advantage of our screening procedure is that we can determine the threshold according to a required false positive rate, whereas other methods usually cut off at ad hoc model sizes. Table 2 compares the effects of these two thresholding approaches. For a targeted false positive rate α, the threshold for MAC_1-F is chosen as the $1 - \alpha$ quantile of MAC_1, simulated under the null hypothesis. We use 500,000 simulations, which actually could be much smaller, to accurately estimate the quantile. Another reason why we use so many simulations is to provide a numerical validation of theoretical results in Theorem 1 and Corollary 1. For the KF, as suggested in Li, Zhong and Zhu (2012b), the threshold is chosen such that the resulting model size is equal to $\lceil n/\log(n) \rceil$ or its multiple. As shown in Table 2, to obtain the same true positives, the KF yields more false positives than MAC_1-F does. By taking different values of α, we show that MAC_1-F controls the false positives reasonably well. Taking $\alpha = 5\%$ as an example, the expected number of false positives is 99.75, with a theoretical standard deviation equal to 9.73, which are close to the values estimated by MAC_1-F.

Table 2. The true/false positives (TP/FP) of the KF and MAC_1-F under different thresholds. The numbers are the mean values from 500 replicates, with standard errors given in parentheses.

Threshold	Example 1	Example 2	Example 3	Example 4	Example 5
$^{\dagger}\alpha$ for MAC_1-F					
5%	5.0(0.14)/ 96.5 (9.8)	4.9(0.3)/ 96.7 (9.5)	5 (0) / 97.3 (9.6)	5.0(0) / 96.7 (9.9)	5(0)/ 96.8(10.0)
1%	4.6(0.6) / 16.8 (4.2)	4.3(0.8)/ 16.9 (4.1)	5.0(0.1)/ 17.4 (4.4)	5 (0) / 17.0 (4.0)	5(0)/ 16.8 (4.6)
0.1%	3.5(1.1) / 3.4 (1.7)	3.1(1.1)/ 3.5 (1.7)	4.9(0.3)/ 3.4 (1.9)	4.9(0.2)/ 3.4 (1.9)	5(0)/ 3.4 (2.0)
d_n for KF					
$\lceil n/\log(n)\rceil$	2.4(1.1) / 40.8 (4.5)	0.8(0.8)/ 43.4 (4.7)	2.7(1.1)/ 41.7 (4.7)	1.7(1.1)/ 42.2 (4.6)	5(0)/ 39.0 (4.2)
$2\lceil n/\log(n)\rceil$	3.1(1.1) / 81.2 (7.3)	1.3(1.0)/ 85.2 (7.8)	3.5(1.2)/ 82.2 (7.6)	2.7(1.1)/ 83.1 (7.6)	5(0)/ 81.1 (7.4)
$3\lceil n/\log(n)\rceil$	3.5(1.1) /125.1(10.7)	1.8(1.1)/126.0(10.4)	4.0(0.9)/125.1(11.4)	3.3(1.1)/124.9(10.7)	5(0)/123.6(10.9)

† The $(1-\alpha)$ quantile of MAC_1 is estimated based on 500,000 simulations.

For the KF, we cannot find an explicit relationship between the chosen model size and the false positives. Thus the accuracy of the resulting model is uncertain.

4.2. Pairwise joint screening

In this section, we evaluate the power of our new pairwise joint screening method using three simple, but representative examples. In Example 6, there is only the joint effect of the condition (J3). In Example 7, we test the power of joint effect screening when we have both a main effect and a joint effect of the condition (J1). In Example 8, we have main and joint effects for conditions (J1) and (J3). Because we focus on joint screening here, we set $p = 300$ and $n = m = 200$ in the simulation. Under this setting, there are 44,850 possible pairwise joint effects, which cannot be handled by a penalized method, in which we assume that the pairwise joint effect is of the form X_1X_2. However, we do not put any assumption on the form of the pairwise joint effect in order to use MAC-F.

Furthermore, we set $\alpha_1 = 0.5\%$ and 5% in $M_1^{\alpha_1}$, as defined by Equation (3.2) for MAC_1-F, $\alpha_{21} = 0.01\%$ in $M_{21}^{\alpha_{21}}$, as defined by Equation (3.5) and $\alpha_{22} = 0.1\%$ in $M_{22}^{\alpha_{22}}$, as defined by Equation (3.6). We use 500,000 simulations to accurately estimate these quantiles. Another reason for using so many simulations is to provide a numerical validation for theoretical results in Theorem 1 and Corollary 2.

Example 6.

- $\log(P(Y=1|X)/P(Y=0|X)) = X_1X_2$.
- $X_j : j = 1, \ldots, p \overset{i.i.d.}{\sim} N(0,1)$.

Table 3. The true/false positives (TP/FP) for Examples 6–8, with $n = m = 200$. The numbers are from 500 replicates, with standard errors given in parentheses.

	Example 6		Example 7		Example 8	
	MAC_1-F†	MAC-F†	MAC_1-F	MAC-F	MAC_1-F	MAC-F
$\alpha_1 = 5\%$						
TP	0.38(0.61)	2(0)	1.2(0.37)	2(0)	1.3(0.51)	4(0)
FP	15(3.8)	4.8(2.1)+9(0.68)§	16(3.9)	4.9(2.2)+9.5(0.67)	16(3.8)	5.4(2.4)+9(0.68)
$\alpha_1 = 0.5\%$						
TP	0.1(0.42)	2(0)	1.1(0.25)	2(0)	1.0(0.19)	4(0)
FP	1.8(1.22)	1.2(1.1)+9(0.67)	1.7(1.21)	1.3(1.0)+8.5(0.69)	1.6(1.22)	1.8(1.3)+9(0.67)

† Quantile of MAC_1 and MAC_2 is estimated based on 500,000 simulations

§ FP of MAC-F is represented by FP of MAC_2-F_2 + FP of MAC_2-F_1

Example 7.

- $\log(P(Y = 1|X)/P(Y = 0|X)) = X_1 + X_1X_2$.
- $X_j : j = 1, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$.

Example 8.

- $\log(P(Y = 1|X)/P(Y = 0|X)) = X_1 + X_1X_2 + X_3X_4$.
- $X_j : j = 1, \ldots, p \overset{i.i.d.}{\sim} N(0, 1)$.

As shown in Table 3, the performance of MAC_1-F for these examples is poor, owing to the existence of the pairwise joint effect. For Example 6, both X_1 and X_2 are marginally undetectable. Thus the true positive of MAC_1-F, as expected, is very small (the TP is less than one). For Example 7, the true variable X_1 is marginally detectable, but X_2 is not. Thus MAC_1-F selects X_1 and misses X_2 very often (the TP is almost one). For Example 8, the true variable X_1 is marginally detectable, but the other three variables are not. Thus MAC_1-F selects X_1 and misses other variables very often (the TP is almost one). However, MAC-F always finds the true variables in these examples, while keeping the false positives under control. Similar conclusions can be drawn from the simulation results for these examples under smaller sample sizes (see Section S2.3 in the **Supplementary Material**).

4.3. Real-data application

In this section, we apply the proposed screening method to an in-house DNA CpG methylation array data set of diabetes patients. The study investigates which genetic and epigenetic factors drive a diabetes patient to develop coronary

heart disease (CHD) or end-stage renal disease (ESRD). The specific scientific goal here is to check which CpG sites are associated with the development of CHD or ESRD among Chinese Type-2 diabetes (T2D) patients. We perform a matched case-control study from the Hong Kong Diabetes Registry, which includes thousands of patients with T2D and prospective follow-up. For control, the patient should have had T2D for ≥ 10 years, but should have neither cardiovascular disease nor chronic kidney disease at both baseline and follow-up. For the CHD case, the patient should have neither CHD nor chronic kidney disease at baseline, but must have developed CHD during follow-up. For the ESRD case, the patient should have neither ESRD nor cardiovascular disease at baseline, but should have developed ESRD during follow-up. Infinium Human Methylation 450K BeadChip is used to measure the methylation levels for each patient. After screening for quality control and phenotype availability, 435 CHD patients and 436 corresponding controls are used for the CHD study, and 363 ESRD patients and 362 corresponding controls are used for the ESRD study. We use 468,034 CpG sites as the candidate variables for screening.

Figure S5 in the **Supplementary Material** shows the K–S statistic values and their corresponding MAC statistic values for all CpG sites. It shows that MAC_1-F found more CpG sites associated with CHD than the KF did. For example, the adjusted p-value of the K–S statistic for the CpG site cg13359998, a reported site contributing to CHD, is one. This means that the KF cannot detect this important site. However, the adjusted p-value of the corresponding MAC statistic is 0.0003. Thus MAC-F successfully found this site. For ESRD, the KF also tends to miss some important CpG sites. The top 40 CpG sites found by the KF and MAC_1-F are provided in the **Supplementary Material**.

We also use our pairwise joint screening procedure to check the joint effect or interaction among the CpG sites. We focus first on detecting pairs of CpG sites under CHD satisfying condition (J1), where the marginally detectable CpG sites are chosen as the top nine CpG sites detected by MAC_1-F (sorted according to the difference in mean methylation level): cg10501210, cg0620273, cg0759483, cg13359998, cg13471990, cg09586924, cg09648727, cg16867657 and cg03139435. The null distribution of $\log(\text{MAC}_2)$ is shown in Figure S6 in the **Supplementary Material**, and is well approximated by a normal distribution. The screening threshold for M_{22}^{α} is set as 52.32, which is the $(1 - 0.05/(40 * 468034))$ quantile of MAC_2 calculated from the normal approximation. The detected CpG sites are provided in the **Supplementary Material**. Interestingly, we find that the CpG site cg19083407, which may strongly interact with CpG site cg09586924, is reported as a potential mediator of genetic association with mRNA expression in

human pancreatic islets (Olsson et al. (2014)).

Next, we focus on detecting pairs of CpG sites under CHD that satisfy condition (J3). Because there are as many as 467994^2 pairs of CpG sites to be detected, we focus on the joint effect of the top 150 from the undetected list of CpG sites. The screening threshold for M_{21}^{α} is set as 64.74, which is the $(1 - 0.1/(467994 * 467993))$ quantile of MAC_2, calculated from the normal approximation. We found that three pairs are significant: cg08985282–cg01287975, cg06655216–cg01287975 and cg26203883–cg08280341. Of these, cg08985282 has been reported to be associated with a mood disorder (Byrne et al. (2013)).

Note that many of our detected CpG sites are novel, and have not been reported as biomedical markers in the current literature. They may provide new insights into studies on diabetes, but independent experiments are needed to validate them. We shall report the validation part of the novel findings in future work.

5. Discussion

We have considered variable screening in high-dimensional binary classification. To this end, we have proposed consistent non-parametric test statistics for the two-sample distribution comparison under different settings. Based on the new statistics, we have proposed new variable screening procedures for ultrahigh-dimensional binary classification, which are much more powerful than existing methods. Importantly, our new pairwise joint screening procedure enables us to find variables with a pairwise joint effect, but no marginal effect. Both the simulation results and the real-data application show the effectiveness and advantages of the proposed methods.

As shown in Theorem 1, our test statistics are consistent under all continuous alternatives when the sample sizes are comparable. Intuitively, the MAC statistics combine the merits of the chi-squared statistic and the K–S statistic. Compared with the chi-squared statistic, the MAC statistics avoid the problem of selecting the number of partitions and constructing the partitions, which are critical, but treated rather arbitrarily in real applications. In comparison with the K–S statistic, the MAC statistics improve the resolution by considering different scales of the neighborhood for each sample point. Our simulation studies showed that MAC_1 outperforms classical methods (K–S, A–D and CvM) in univariate cases, and MAC_3 outperforms **ZZZ** in high-dimensional cases. However, in univariate cases, where the parametric assumption on the alternatives of **ZZZ** holds, **ZZZ** performed slightly better than MAC_1, owing to its parametric ef-

ficiency. The improved performance of MAC for the two-sample distribution comparison shows its potential for variable screening in high-dimensional binary classification.

As illustrated in our examples, the K–S test is powerless in many cases, which leads to the low true positives of the KF. Another drawback of the KF is that there is no way to determine the number of variables that should be included in the model. Our MAC-based screening procedure, MAC_1-F, overcomes these drawbacks, owing to its consistent screening property rather than the sure screening property. The threshold for screening variables can be set according to the Type-I error in practice. This is another advantage of MAC_1-F over the KF. However, the power gain comes at the cost of computational efficiency.

Considering that MAC_1-F screens variables according to marginal effect, we proposed a new joint screening procedure based on MAC_2 and MAC_1, called MAC-F. As shown in our examples, marginal screening methods, such as MAC_1-F, may miss some important variables that are not marginally detectable. MAC-F enables us to find these kinds of variables. If there are p input variables, we shall compute $\binom{p}{2}$ MAC_2 values, which is time-consuming when p is large. Although this problem cannot be avoided by other screening methods, parallel techniques can be used to reduce the computation time. Furthermore, we only consider second-order joint screening in this work. It is still interesting to develop higher-order joint screening methods based on MAC_3, which is left to future work.

Although the MAC statistics are powerful for both the two-sample distribution comparison and variable screening in binary classification, their computation may be time-consuming. Note that the computation time of MAC_1 and MAC_2 are both $O(nm(n+m))$, whereas the computation time of MAC_3 is $O(\binom{d}{2}nm(n+m))$. Thus, MAC_3 has the limitation of higher computation complexity when d becomes large. As an example, the computation time of MAC_3 for Example 9 ($d = 5$) in Section S1.2 in the **Supplementary Material**, with sample size $n = m = 200$, is 3.6 seconds on a Windows7 PC with an Intel Core i7-2600 3.4GHz processor. There are several possible approaches to reduce the computing time of the MAC statistics without significant sacrifice of accuracy. For example, in Equations (2.1)-(2.3), $\max\{T_k(\mathbf{x}_i, \mathbf{y}_j), T_k(\mathbf{y}_j, \mathbf{x}_i)\}$ can be replaced by $T_k(\mathbf{x}_i, \mathbf{y}_j)$, which results in only a minor decrease of accuracy, but halves of the computing time. One can also compute the MAC statistics on a sub-sample instead of scanning all sample points with all sample-based resolutions. In addition, the MAC statistics can be computed in a parallel manner by computing the local statistics separately, which would save a lot of time. However, additional research is needed to extend the idea behind MAC_2 in a more efficient way to

cases with $d > 2$, rather than using the naive exhaustive bi-partition approach currently adopted by MAC_3. When the sample sizes are not big, but the increasing dimension causes a computational problem, gCvM may be a good alternative, as long as we are aware of its weakness.

Supplementary Material

The online Supplementary Material provides additional simulation results and proofs of the theoretical results.

Acknowledgments

We would like to thank the associate editor and two referees for their constructive comments. In addition, we thank the authors of Zhou et al. (2017) for sharing their code with us. This research was partially supported by The Research Grant Council of the Hong Kong Special Administrative Region, China (Theme-based Research Scheme T12-402/13N; General Research Fund No. 14203915, No. 14173817, No. 15301218, and No. 15303319), and The National Natural Science Foundation of China (No. 11901517 and No. 11771366).

References

Bera, A. K., Ghosh, A. and Xiao, Z. J. (2013). A smooth test for the equality of distributions. *Econometric Theory* **29**, 419-446.

Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123**, 160-171.

Byrne, E. M., Carrillo-Roa, T., Henders, A. K., Bowdler, L., McRae, A. F., Heath, A. C. et al. (2013). Monozygotic twins affected with major depressive disorder have greater variance in methylation than their unaffected co-twin. *Translational Psychiatry* **3**, e269.

Chen, H. and Friedman, J. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association* **112**, 397-409.

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.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests. *The Annals of Mathematical Statistics* **28**, 823-838.

Eubank, R. L. and LaRiccia, V. N. (1992). Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *The Annals of Statistics* **20**, 2071-2086.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association* **91**, 674–688.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics* **36**, 2605-2637.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849-911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **70**, 3567–3604.

Fan, J., Feng, Y. and Song, R. (2012). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106**, 544-557.

Friedman, J. and Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* **7**, 697-717.

Hall, P. and Miller, H. (2012). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics* **18**, 533-550.

Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89**, 359-374.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics* **16**, 772–783.

Huang, Q. and Zhu, Y. (2016) Model-free sure screening via maximum correlation. *Journal of Multivariate Analysis* **148**, 89–106.

Kim, I., Balakrishnan, S. and Wasserman, L. (2020) Robust Multivariate Nonparametric Tests via Projection-Pursuit. *The The Annals of Statistics. arXiv preprint arXiv:1803.00715*. To appear.

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

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

Ludwig, B. and Carsten, F. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis* **88**, 190-206.

Ludwig, B. and Carsten, F. (2010). Rigid motion invariant two-sample tests. *Statistica Sinica* **20**, 1333-1361

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **1**, 229–234.

Neyman, J.(1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 149–199.

Olsson, A., Volkov, P., Bacos, K., Dayeh, T., Hall, E., Nilsson, E. A. et al. (2014). Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets. *PLoS Genetics* **10**, e1004735.

Rosenbaum, P. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 515–530.

Schilling, M. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association* **81**, 799–806.

Székely, G. and Rizzo, M. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143**, 1249–1272.

Thas, O. (2010). *Comparing distributions*. Springer-Verlag, New York.

Weiss, L. (1960). Two-sample tests for multivariate distributions. *The Annals of Mathematical Statistics* **31**, 159–164.

Zhou, W., Zheng, C. and Zhang, Z (2017). Two-sample smooth tests for the equality of distri-

butions. *Bernoulli* **23**, 951–989.

Hangjin Jiang
Center for Data Science, Zhejiang University, Hangzhou 310058, China.
E-mail: jianghj@zju.edu.cn
Xingqiu Zhao
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
E-mail: xingqiu.zhao@polyu.edu.hk
Ronald C.W. Ma
Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.
E-mail: rcwma@cuhk.edu.hk
Xiaodan Fan
Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.
E-mail: xfan@cuhk.edu.hk

(Received September 2018; accepted June 2020)