Consistent Screening Procedures in High-dimensional Binary Classification

Hangjin Jiang[†,††], Xingqiu Zhao[‡], Ronald C.W. Ma[*], Xiaodan Fan[††,1]

[†] *Center for Data Science, Zhejiang University*

[‡]*Department of Applied Mathematics, The Hong Kong Polytechnic University*

[*]*Department of Medicine and Therapeutics, The Chinese University of Hong Kong*

[††]*Department of Statistics, The Chinese University of Hong Kong*

Supplementary Material

Section S1 presents further simulation results to show the performance of MAC statistics for distribution comparison under varies of settings.

Section S2 presents further simulation results to show the performance of MAC_1-F for variable screening on extra examples where KF performs very well, and simulation results under smaller sample sizes for examples in the main paper.

Sections S3 presents additional results on empirical distribution of MAC and real data example.

Sections S4 contains all the proofs of theoretical results in the main paper.

[1] Corresponding Author: xfan@cuhk.edu.hk

S1 Two-Sample Distribution Comparison

In this section, additional simulations are conducted to explore the finite sample performance of our new tests and compare them with other methods. The statistical power of each method on each example is estimated from 2000 independent replications under significance level 0.05.

S1.1 Univariate Case

In this section, we evaluate the performance of MAC_1 for two-sample distribution comparison in one-dimensional case through comparing it with the K-S test, the A-D statistic, CvM and the latest method proposed in Zhou et al. [2017] (denoted as **ZZZ**) based on the following five examples with either bumps or high-frequency components. These examples are adopted from Zhou et al. [2017], which were partially collected from Fan [1996] and Heyde [2010]. Example 1 corresponds to the case where the differences between two distributions are local bumps with various widths and magnitudes. Example 2 illustrates the effect of global features with different frequencies. Example 3 is designed to test distributions different by an oscillating term of intermediate frequency in the log scale. Example 4 and 5 aim to cover the high-frequency alternatives.

Example 1. $X \sim F = \text{uniform}(-1, 1)$ versus $Y \sim G = G_c$ with density $g_c(x) = \{0.5 + 2x\frac{c-|x|}{c^2}I(|x| < c)\}I(|x| \leq 1)$, $(0 \leq c \leq 1)$.

Example 2. $X \sim F = \text{uniform}(-1, 1)$ versus $Y \sim G = G_c$ with density $g_c(x) = \{0.5 + 0.5\sin(2\pi c x)\}I(|x| \leq 1)$, $(0.5 \leq c \leq 5)$.

Example 3. $X \sim F = \text{lognormal}(0, 1)$ with density $f(x) = (2\pi)^{-1/2}x^{-1}\exp\{-(\log x)^2/2\}$ versus $Y \sim G = G_c$ with density $g_c(x) = f(x)(1 + c\sin(2\pi \log x))$, $(-1 \leq c \leq 1)$

Example 4. $X \sim F = \text{uniform}(0, 1)$ versus $Y \sim G = G_c$ with density $g_c(x) = \exp\{c\sin(5\pi x)\}I(0 \leq x \leq 1)$, $(0 \leq c \leq 2)$.

Example 5. $X \sim F = \text{uniform}(0, 1)$ versus $Y \sim G = G_c$ with density $g_c(x) = \{1 + c\cos(5\pi x)\}I(0 \leq x \leq 1)$, $(0 \leq c \leq 2)$.

The empirical power of these methods for Example 1-5 under various settings are presented in Figure S1 and Figure S2. Figure S1 focuses on comparing these five methods under the same setting across different examples. Specifically, in Example 1, our new test performed the best. For Example 2, **ZZZ** performed better in the low frequency settings when sample sizes are small, but its advantage disappeared with the increase of sample sizes; But, our new test always performed the best in the high frequency settings regardless of the sample sizes. For Examples 3-5, where the parametric assumption of **ZZZ** held, **ZZZ** slightly outperformed our new test, but the performance difference reduces with the increase of the sample sizes. This is reasonable since parametric methods, compared with nonparametric methods, have the advantage of requiring a smaller sample size to obtain the same power on the data satisfying their assumptions. However, in cases where their parametric assumption fails (Example 1 and 2), our new test usually outperforms **ZZZ**. The K-S test, A-D and CvM have a lower power in these examples, which means they are unable to detect densities with bumps or high frequent components.

Figure S2 shows the performance of each method under different sample sizes. As expected, the power of our new method increases with the sample sizes. However, this is not the case for other methods. For example, the power of **ZZZ** in Example 2, comparing with that of MAC_1, changed a little when the sample sizes are increased from (180, 150) to (200, 200). This may be due to the restrictive assumption of **ZZZ** on the alternative density. Similarly, the power of A-D, K-S and CvM in Example 3 stay at a very low level under different sample sizes due to their weakness in detecting densities with high frequent components.

S1.2 Multi-dimensional Case

In this section, we compare the finite sample performance of MAC_3 with that of **ZZZ** and gCvM [Kim et al., 2020] in multi-dimensional cases under the setting $n = 180$ and $m = 160$ based on examples taken from Zhou et al. [2017].

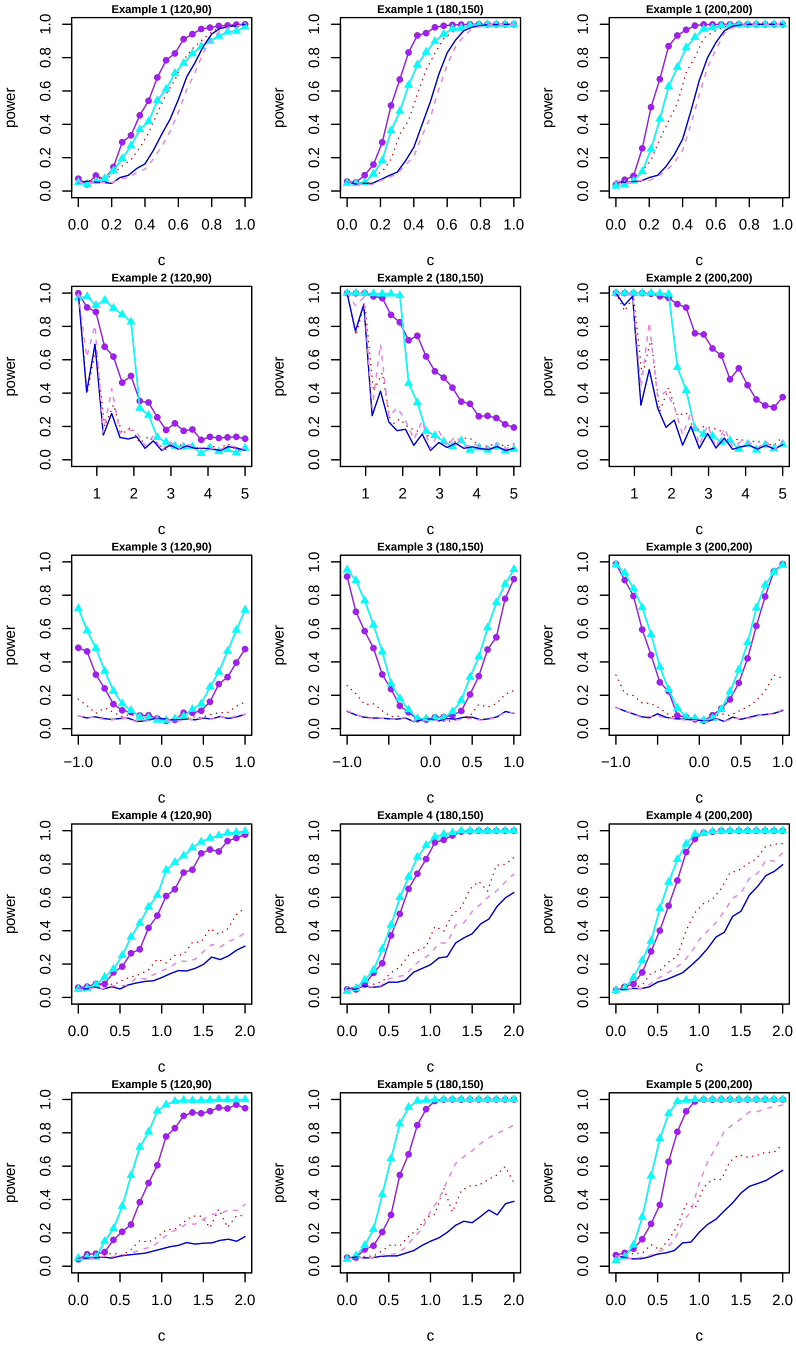

Figure S1: Power of K-S (········), A-D (- - - -), CvM (——), MAC$_1$ (—●—), and **ZZZ** (—▲—) for Example 1-5 based on 2000 independent replications under significance level 0.05. In this subplot matrix, different rows correspond to different examples and different columns correspond to different sample size settings.

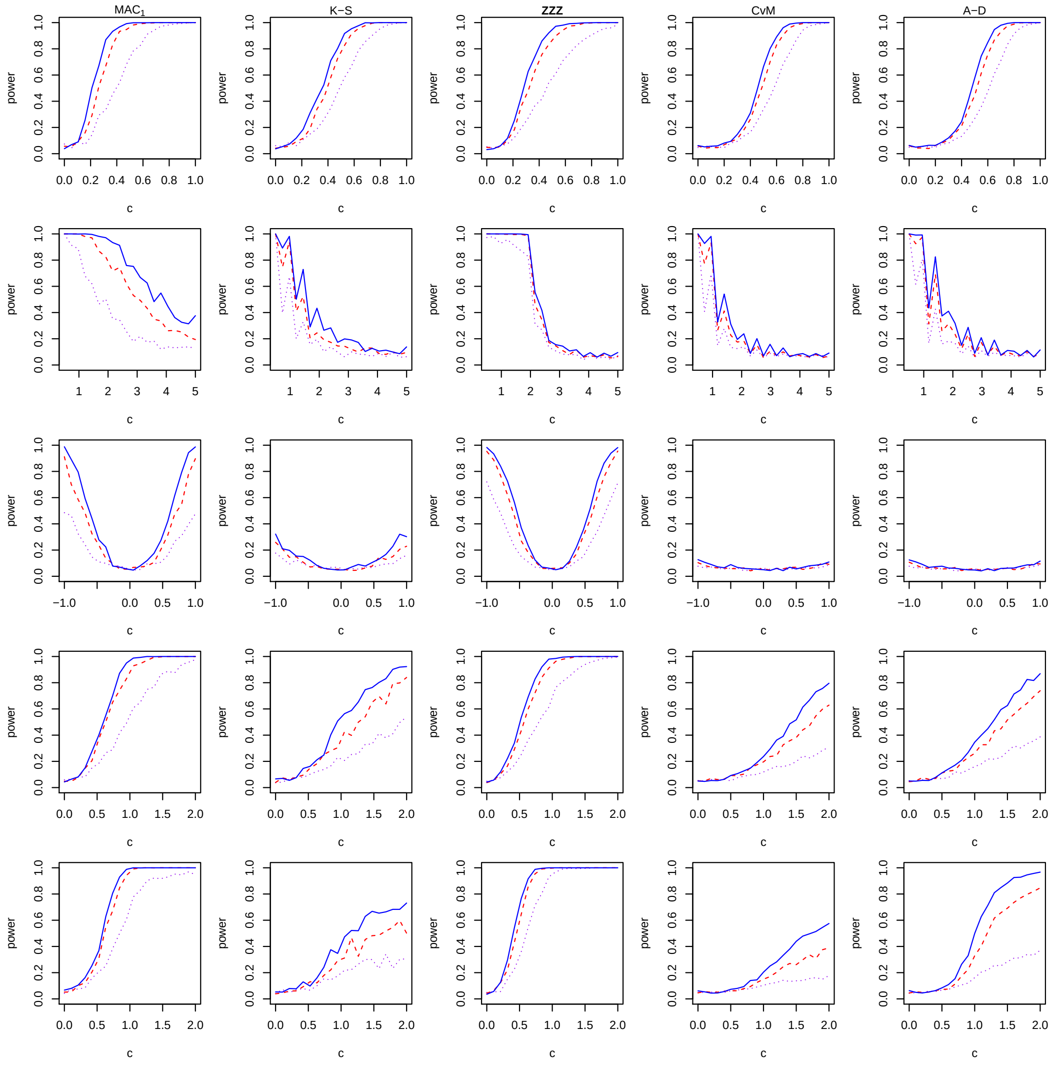

Figure S2: Power of MAC$_1$, K-S, **ZZZ**, CvM and A-D (corresponding to Column 1-5 respectively) under different sample sizes (⋯⋯for n=120 and m=90; - - - -for n=180 and m=150; ——for n=200 and m=200) for Example 1-5 (corresponding to Row 1-5 respectively) based on 2000 independent replications under significance level 0.05

Example 6 and 7 consider two distributions with the same dependence structure but different marginal distributions. Specifically, the marginal distributions in Example 6 are different in local bumps with various widths and magnitudes. In Example 7, the marginal distributions are different in global features of a high frequency and different magnitude. Example 8 and 9 are two five-dimensional examples constructed from Gaussian distribution and t distribution, respectively. In either example, the two distributions differ in the first two dimensions in terms of both marginal distribution and dependence structure.

Example 6. $\mathbf{X} = (X_1, X_2, X_3)', X_1, X_2 \sim_{i.i.d.}$ Uniform$(-1, 1), X_3 = 0.3X_1+0.7X_2$ versus $\mathbf{Y} = (Y_1, Y_2, Y_3)', Y_1, Y_2 \sim_{i.i.d.} g_c(x) = \{0.5+2x\frac{c-|x|}{c^2}I(|x| < c)\}I(|x| \leq 1), (0 \leq c \leq 1)$, and $Y_3 = 0.3Y_1 + 0.7Y_2$.

Example 7. $\mathbf{X} = (X_1, X_2, X_3)', X_1, X_2 \sim_{i.i.d.}$ Uniform$(0, 1), X_3 = 0.3X_1+ 0.7X_2$ versus $\mathbf{Y} = (Y_1, Y_2, Y_3)', Y_1, Y_2 \sim_{i.i.d.} g_c(x) = \exp\{c \sin(5\pi x)\}I(0 \leq x \leq 1), (0 \leq c \leq 2)$, and $Y_3 = 0.3Y_1 + 0.7Y_2$.

Example 8. $\mathbf{X} \sim N(0, I_5)$ versus $\mathbf{Y} = A\mathbf{Z}, \mathbf{Z} \sim N(0, I_5)$, where

$$A = \begin{pmatrix} A_0 & 0 \\ 0 & I_3 \end{pmatrix}, \; A_0 = \begin{pmatrix} \sqrt{1-c} & \sqrt{c} \\ \sqrt{c} & \sqrt{1-c} \end{pmatrix}, \; (0 \leq c \leq 0.5).$$

Example 9. $\mathbf{X} \sim t_4(0, I_5)$ versus $\mathbf{Y} = A\mathbf{Z}, \mathbf{Z} \sim t_4(0, I_5)$, where

$$A = \begin{pmatrix} A_0 & 0 \\ 0 & I_3 \end{pmatrix}, \ A_0 = \begin{pmatrix} \sqrt{1-c} & \sqrt{c} \\ \sqrt{c} & \sqrt{1-c} \end{pmatrix}, \ (0 \le c \le 0.5).$$

As shown in Figure S3, MAC$_3$ outperformed **ZZZ** and gCvM in all these four examples, and the power improvement of MAC$_3$ over **ZZZ** is dramatic in Example 8 and 9, the 5-dimensional case. The lower power of **ZZZ** may due to the difficulty from the corresponding optimization problem. gCvM performs the worst in Example 6 and 7 due to bumps and high-frequency components, and also in Example 8 and 9 due to its insensitiveness to local differences.

S2 Additional Simulation Results for Screening Procedures

S2.1 Example 4.1-4.5 with smaller sample sizes

In this section, we present simulation results for Example 4.1-4.5 under smaller sample sizes. Table S1 shows the results for the case where sample sizes are $n = m = 50$, Table S2 corresponds to the case where sample sizes are $n = m = 100$, and Table S3 is for the case where sample sizes are $n = m = 150$. MAC$_1$-F significantly outperformed KF in all cases.

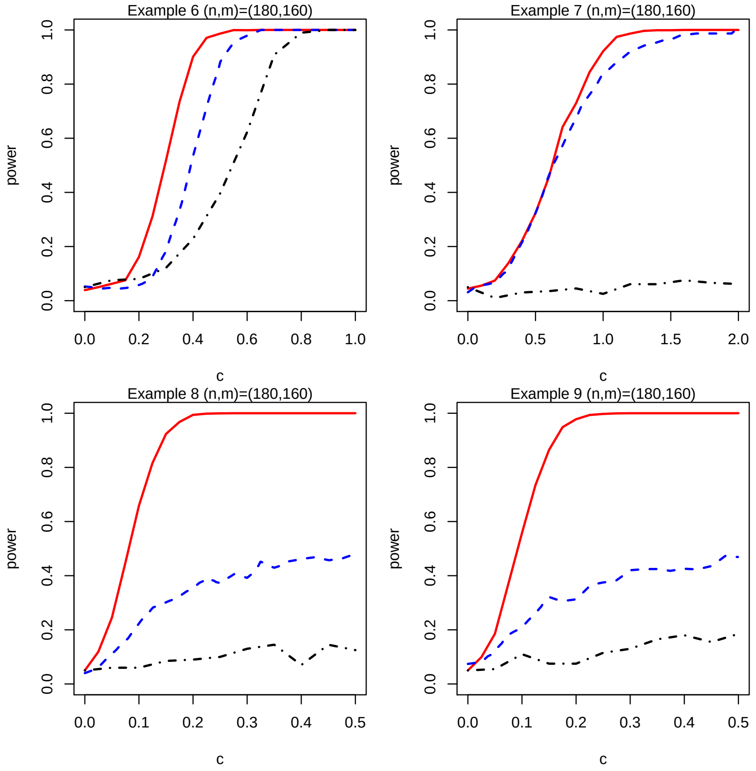

Figure S3: Power of MAC$_3$ (——), **ZZZ** (- - - -) and gCvM (-·-·-) for Example 6-9 based on 2000 independent replications under significance level 0.05.

Table S1: Smallest model size required to contain all the true variables for sample sizes $n = m = 50$. The numbers are medians from 500 replicates with the standard errors of these median estimates (estimated by Bootstrap) given in parentheses.

Method	Example 4.1	Example 4.2	Example 4.3	Example 4.4	Example 4.5
KF	820(66)	1500(28)	1100(6.9)	1100(4.1)	16(0.33)
MAC$_1$-F	570(20)	1100(31)	560(19)	530(27)	5(0)

Table S2: Smallest model size required to contain all the true variables for sample sizes $n = m = 100$. The numbers are medians from 500 replicates with the standard errors of these median estimates (estimated by Bootstrap) given in parentheses.

Method	Example 4.1	Example 4.2	Example 4.3	Example 4.4	Example 4.5
KF	310(5.8)	1200(5.2)	580(26)	730(6.8)	5(0)
MAC$_1$-F	78(3.9)	400(27)	63(4.5)	43(2)	5(0)

Table S3: Smallest model size required to contain all the true variables for sample sizes $n = m = 150$. The numbers are medians from 500 replicates with the standard errors of these median estimates (estimated by Bootstrap) given in parentheses.

Method	Example 4.1	Example 4.2	Example 4.3	Example 4.4	Example 4.5
KF	120(5.7)	900(7.2)	360(8.3)	460(5.9)	5(0)
MAC$_1$-F	16(0.47)	96(6.6)	9(0.35)	7(0.46)	5(0)

S2.2 Extra Examples for Comparing MAC$_1$-F with KF

In this section, we show additional simulation results based on examples taken from Mai and Zou [2013]. We set $p = 2000$ and $n = m = 200$ in our simulation. For each example, 500 independent experiments are performed to evaluate the performance of these two methods. Results are summarized in Table S4, which shows that these two methods work equally well.

Example B1

- $X_j|Y = 1 \sim N(1.922, 1)$, $X_j|Y = 0 \sim N(0, 1)$, $j = 1, \cdots, 8$
- $X_j : j = 9, \cdots, p \sim_{i.i.d.} N(0, 1)$

Example B2

- $\log \frac{P(Y=1|X)}{P(Y=-1|X)} = -3 + 2X_1 + 2X_2 + 2X_3 + 3\sin(X_4) + 4X_5^2$
- $X_j : j = 6, \cdots, p \sim_{i.i.d.} N(0, 1)$

Example B3

- $X|Y = 1 \sim N(\mu, \Sigma)$, $X|Y = 0 \sim N(0, \Sigma)$ where $\mu = \Sigma\beta, \beta = -0.41 \times 1_8, \Sigma = (\sigma_{ij})_{8\times 8}, \sigma = 0.8^{|i-j|}$, $1_8 \in R^8$ is a vector whose elements are all 1.
- $X_j : j = 9, \cdots, p \sim_{i.i.d.} N(0, 1)$

Example B4

Table S4: Smallest model size required to contain all the true variables for Example B1-B6. The numbers are medians from 500 replicates with standard errors (estimated by Bootstrap) given in parentheses.

Method	Example B1	Example B2	Example B3	Example B4	Example B5	Example B6
KF	8(0)	5(0)	8(0)	8(0)	4(0)	4(0)
MAC-F	8(0)	5(0)	8(0)	8(0)	4(0)	4(0)

- $W|Y$ follows the model in Example B3, let $X_j = e^{2W_j}$ for $j = 1, \cdots, 8$.
- $X_j : j = 9, \cdots, p \sim_{i.i.d.} N(0, 1)$

Example B5

- $X|Y = 1 \sim N(\mu, \Sigma), X|Y = 0 \sim N(0, \Sigma)$ where $\mu = 0.63\times(1, -1, -1, 1), \sigma_{ij} = 0.8, i \neq j$.
- $X_j : j = 5, \cdots, p \sim_{i.i.d.} N(0, 1)$

Example B6

- $W|Y$ follows the model in Example B5, let $X_j = e^{2W_j}$ for $j = 1, \cdots, 4$.
- $X_j : j = 5, \cdots, p \sim_{i.i.d.} N(0, 1)$

S2.3 Example 4.6-4.8 with smaller sample sizes

In this section, we present simulation results for Example 4.6-4.8 under smaller sample sizes. Table S5 presents the results for the case with sample sizes $n =$

Table S5: The True/False positive (TP/FP) for Example 4.6-4.8 with $n = m = 100$. The numbers are from 500 replicates with standard errors given in parentheses.

	Example 4.6		Example 4.7		Example 4.8	
	MAC$_1$-F[†]	MAC-F[‡]	MAC$_1$-F	MAC-F	MAC$_1$-F	MAC-F
$\alpha_1 = 5\%$						
TP	0.33(0.55)	2(0)	1.17(0.38)	2(0)	1.21(0.46)	4(0)
FP	15.5(3.5)	4.7(2.2)+9.5(0.67)[§]	16(3.8)	4.8(2.1)+9.6(0.68)	16(3.7)	5.3(2.5)+9.5(0.67)
$\alpha_1 = 0.5\%$						
TP	0.09(0.21)	2(0)	1.03(0.18)	2(0)	1.0(0.12)	4(0)
FP	1.6(1.15)	1.3(1.0)+9.4(0.66)	1.65(1.2)	1.25(1.1)+9(0.67)	1.65(1.21)	1.7(1.2)+9(0.68)

[†] Quantile of MAC$_1$ is estimated based on 500000 simulations

[‡] Quantile of MAC$_2$ is estimated based on 500000 simulations

[§] FP of MAC-F is represented by FP of MAC$_2$-F$_2$ + FP of MAC$_2$-F$_1$

$m = 100$, and Table S6 presents the results for the case with sample sizes $n = m = 150$. It shows that MAC-F performed well while MAC$_1$-F may miss some important variables.

S3 Additional details on real data application

In this section, we show additional results from the real data application. Figure S4 shows the empirical distribution of MAC$_1$ and MAC$_2$. Figure S5 shows the K-S values and MAC$_1$ values of CpG sites for CHD and ESRD. Figure S6 shows the empirical distribution of $\log(\text{MAC}_2)$ under H_0.

Table S6: The True/False positive (TP/FP) for Example 4.6-4.8 with $n = m = 150$. The numbers are from 500 replicates with standard errors given in parentheses.

	Example 4.6		Example 4.7		Example 4.8	
	MAC$_1$-F†	MAC-F‡	MAC$_1$-F	MAC-F	MAC$_1$-F	MAC-F
$\alpha_1 = 5\%$						
TP	0.37(0.62)	2(0)	1.21(0.41)	2(0)	1.26(0.46)	4(0)
FP	15.5(3.7)	4.9(2.0)+9.2(0.67)§	16(3.7)	5(2.1)+9.4(0.68)	15.7(3.7)	5.2(2.1)+9.3(0.62)
$\alpha_1 = 0.5\%$						
TP	0.1(0.32)	2(0)	1.06(0.23)	2(0)	1.0(0.21)	4(0)
FP	1.7(1.21)	1.3(1.2)+9.1(0.66)	1.8(1.21)	1.25(1.1)+9(0.68)	1.5(1.21)	1.75(1.3)+9.2(0.67)

† Quantile of MAC$_1$ is estimated based on 500000 simulations

‡ Quantile of MAC$_2$ is estimated based on 500000 simulations

§ FP of MAC-F is represented by FP of MAC$_2$-F$_2$ + FP of MAC$_2$-F$_1$

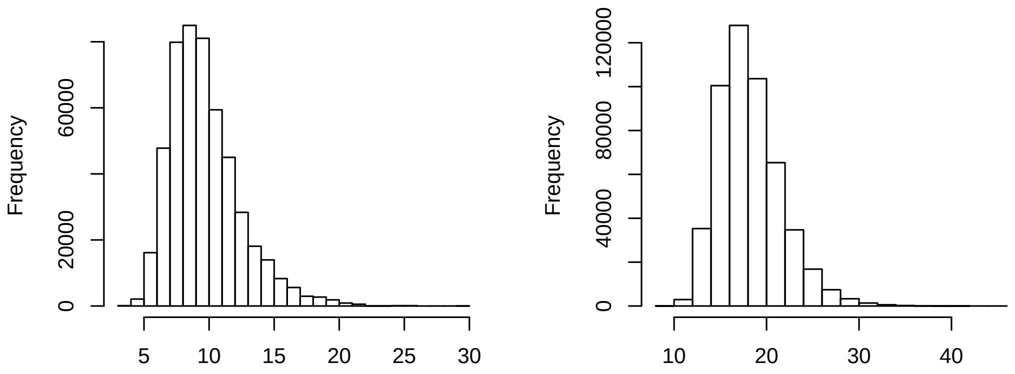

Figure S4: Empirical distribution of MAC$_1$ (Left) and MAC$_2$ (Right) under the null hypothesis based on 500,000 independent experiments with $n = m = 200$.

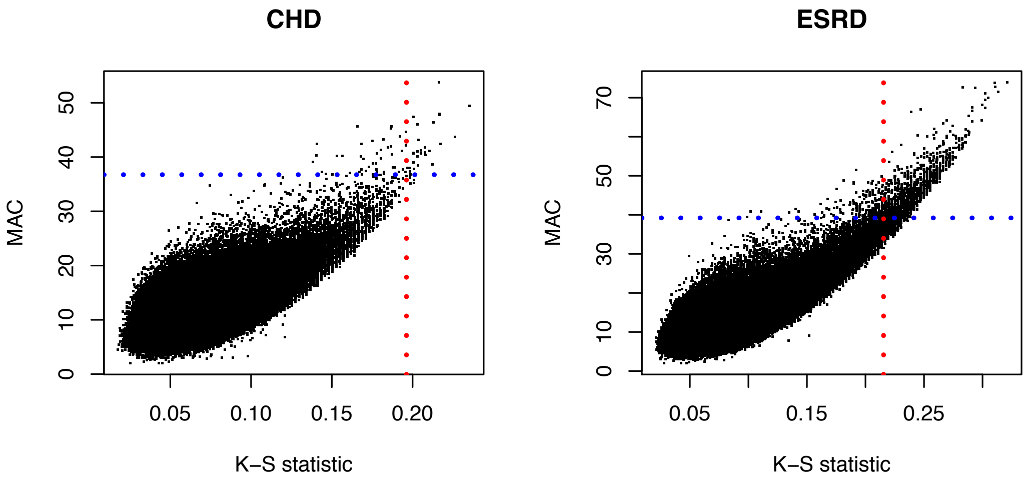

Figure S5: K-S values versus MAC values of all CpG sites for CHD and ESRD. The vertical and horizontal dotted lines corresponds to the screening threshold of MAC-F and KF, respectively.

S4 Proof of Theoretical Results

Firstly, we introduce some notations used in our proofs. Let $X \to_d Y$ denote the convergence of X to Y in distribution, $X \to_p Y$ the convergence of X to Y in probability, $X \to_{a.s.} Y$ denotes that X converges almost surely to Y, and $X =_d Y$ denotes that X and Y have the same distribution.

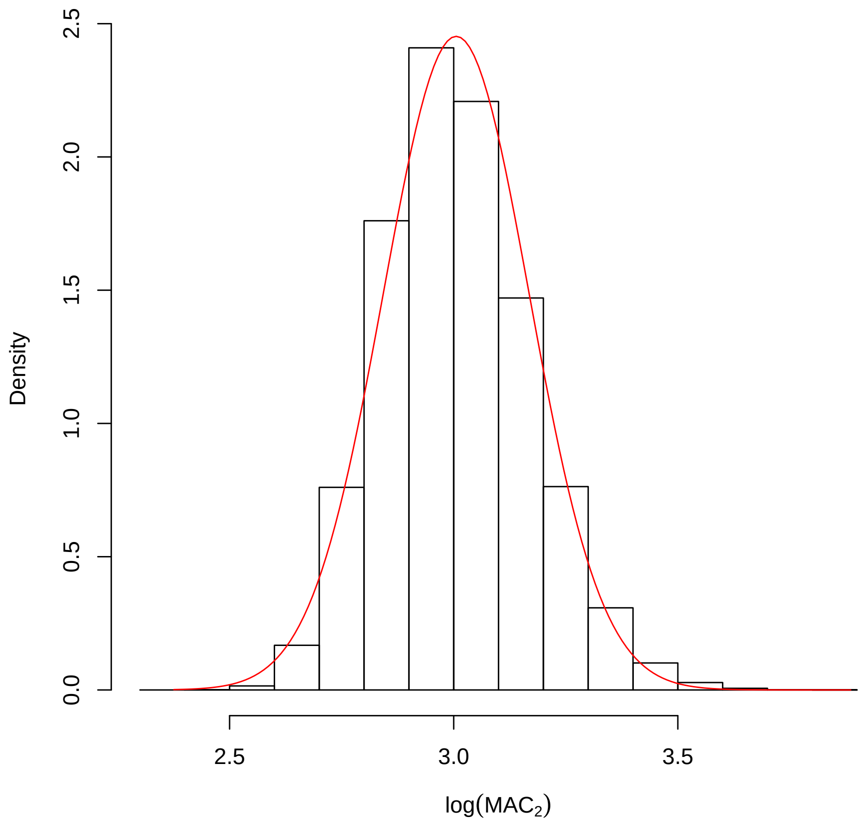

Figure S6: Empirical distribution of $\log(\text{MAC}_2)$ under H_0 with $n = 435, m = 436$ from 1000,000 simulations with its normal approximation. The red lines is the density function of $N(3.006, 0.163)$.

Lemma S.1. *Suppose that* $\boldsymbol{U}_m = (U_{m,1}, \cdots, U_{m,k})$ *and* $\boldsymbol{V}_n = (V_{n,1}, \cdots, V_{n,k})$ *are independently sampled from multinominal distributions with parameters* $(m, a_1, \cdots, a_k)$ *and* $(n, b_1, \cdots, b_k)$*, respectively. Define the column vectors* $\boldsymbol{a} = (a_1, \cdots, a_k)$ *and* $\boldsymbol{b} = (b_1, \cdots, b_k)$*. Under the null hypothesis* H_0*:* $\boldsymbol{a} = \boldsymbol{b}$*, we have* $T = \sum_{i=1}^{k} \frac{(U_{m,i} - m\hat{c}_i)^2}{m\hat{c}_i} + \frac{(V_{n,i} - n\hat{c}_i)^2}{n\hat{c}_i} \to \chi^2_{k-1}$*, where* $\hat{\boldsymbol{c}} = \frac{\boldsymbol{U}_m + \boldsymbol{V}_n}{n+m}$ *be the estimate of* $\boldsymbol{a}$ *under* H_0*.*

Proof. Let $\{X_t\}_{t=1}^m$ and $\{Y_t\}_{t=1}^n$ be the observed category series of the count vectors $\mathbf{U}_m$ and $\mathbf{V}_n$, respectively. So we have $P(X_t = i) = a_i$ and $P(Y_t = i) = b_i$. Let $N = n + m$, we define

$$X_i^* = (X_{1,i}^*, \cdots, X_{N,i}^*) = (\delta_{i,X_1}, \cdots, \delta_{i,X_m}, 0, \cdots, 0)$$

$$Y_i^* = (Y_{1,i}^*, \cdots, Y_{N,i}^*) = (0, \cdots, 0, \delta_{i,Y_1}, \cdots, \delta_{i,Y_n})$$

where $\delta_{i,j} = I(i = j)$.

Firstly, we have $U_{m,i} - m\hat{c}_i = \frac{nU_{m,i} - mV_{n,i}}{N}$, and $V_{n,i} - n\hat{c}_i = \frac{-nU_{m,i} + mV_{n,i}}{N}$.

Thus,

$$T = \sum_{i=1}^{k} \frac{(U_{m,i} - m\hat{c}_i)^2}{m\hat{c}_i} + \frac{(V_{n,i} - n\hat{c}_i)^2}{n\hat{c}_i} = \sum_{i=1}^{k} \frac{(nU_{m,i} - mV_{n,i})^2}{nmN\hat{c}_i}.$$

Let $Z_i = nU_{m,i} - mV_{n,i} = n\sum_{k=1}^{N} X_{k,i}^* - m\sum_{k=1}^{N} Y_{k,i}^*$, under H_0, we have $E(Z_i) = 0, \text{Var}(Z_i) = nm(n+m)a_i(1-a_i), \text{Cov}(Z_i, Z_j) = -nm(n+m)a_ia_j$. So, by the multivariate central limit theorem, we have

$$\frac{1}{\sqrt{nmN}}\mathbf{Z} = \frac{n\mathbf{U}_m - m\mathbf{V}_n}{\sqrt{nmN}} \to_d N(0, \Sigma),$$

where $\sigma_{i,j}$, the (i,j)-th element of Σ, is equal to $a_i(\delta_{ij} - a_j)$. Since $\hat{\mathbf{c}} \to_p \mathbf{a}$, we have

$$\frac{n\mathbf{U}_m - m\mathbf{V}_n}{\sqrt{nmN\hat{\mathbf{c}}}} \to_d N(0, I_k - \sqrt{\mathbf{a}}\sqrt{\mathbf{a}}')$$

Thus, we have $T \to \chi^2_{k-1}$, as $n, m \to +\infty$. □

S4.1 Proof of Theorem 1

Lemma S.2. *Suppose that $X_1, X_2, \cdots, X_n, X$ are i.i.d. sub-exponential with parameters (v, b), i.e., $E(e^{\lambda(X-\mu)}) \leq e^{v^2\lambda^2/2}$ for all $|\lambda| \leq 1/b$, where $\mu = E(X)$, then*

$$\max_{1\leq i\leq n} X_i \leq \max\{2b\log(n), v\sqrt{2\log(n)}\} + \mu$$

holds with probability going to 1 as $n \to +\infty$.

Proof. If we have

$$P(X \geq \mu + t) \leq \begin{cases} e^{-\frac{t^2}{2v^2}}, & 0 \leq t \leq v^2/b \\ e^{-\frac{t}{2b}}, & \text{otherwise} \end{cases} \tag{A.1}$$

then $P(\max_{1\leq i\leq n} X_i \geq \mu + t) \leq nP(X \geq \mu + t)$. So, if $t_1 = v\sqrt{2\log(n^\alpha)} \leq v^2/b$ for any $\alpha > 1$, we have $nP(X \geq \mu + t_1) \leq n^{1-\alpha}$. If $t_2 = 2b\log(n^\alpha) > v^2/b$ for any $\alpha > 1$, we have $nP(X \geq \mu + t_2) \leq n^{1-\alpha}$.

Thus, we have $\max_{1\leq i\leq n} X_i < \mu + \max\{\inf_{\alpha>1} t_2, \inf_{\alpha>1} t_1\} = \mu + \max\{2b\log(n), v\sqrt{2\log(n)}\}$ with probability going to 1.

To prove $(A.1)$, basic analyses based on the Chernoff inequality, i.e., $P(X - \mu \geq t) \leq e^{-\lambda t} E e^{\lambda(X-\mu)}$, can do the work. □

Lemma S.3. *Suppose X is χ_n^2 distributed, then X is sub-exponential with parameters $(2\sqrt{n}, 4)$.*

Proof. It is easy to show that $Y \sim \chi_1^2$ is sub-exponential with parameters $(2, 4)$, and $X = \sum_{i=1}^n Y_i$ where $Y_i \sim_{i.i.d.} \chi_1^2$. Thus $Ee^{\lambda(X-n)} = E(e^{\lambda(\sum_{i=1}^n (Y_i - 1))}) \leq e^{4n\lambda^2/2}$ for $|\lambda| \leq 1/4$, and the result follows. □

Corollary S.1. *Suppose that $Z, Z_1, \cdots, Z_n$ i.i.d $\sim \chi_k^2$, we have $P(\max_{1 \leq i \leq n} Z_i \leq 8\log(n) + k) \to 1$ as $n \to +\infty$.*

Proof. By Lemma S.3, Z is sub-exponential with parameters $(2\sqrt{k}, 4)$. By Lemma S.2, we have $\max_{1 \leq i \leq n} Z_i \leq 8\log(n) + k$ with probability going to 1 as $n \to +\infty$. □

Lemma S.4. *Let $\rho(X, Y)$ be the correlation between random variables X and Y. Suppose that $Z_1, Z_2, \cdots, Z_n$ are random variables such that $\rho(Z_i, Z_j) \geq 0$, for any $1 \leq i < j \leq n$, and $Z_1^*, Z_2^*, \cdots, Z_n^*$ are independent random variables such that $Z_i =_d Z_i^*$, for any $1 \leq i \leq n$, we have $P(\max_{1 \leq i \leq n} Z_i > t) \leq P(\max_{1 \leq i \leq n} Z_i^* > t)$ for any $t \in R$.*

Proof. Define $A_i = \{Z_i > t\}$ and $A_i^* = \{Z_i^* > t\}$ for any i for convenience, and we have $P(A_i) = P(A_i^*)$ by assumption. The proof goes through induction.

When $n = 1$, it is easy to see the conclusion holds.

Now, we consider the case $n = 2$. We have $P(\max\{Z_1, Z_2\} > t) = P(A_1) + P(A_2) - P(A_1A_2)$, and $P(\max\{Z_1^*, Z_2^*\} > t) = P(A_1^*) + P(A_2^*) - P(A_1^*)P(A_2^*) = P(A_1) + P(A_2) - P(A_1)P(A_2)$. Now, we consider two random variables I_{A_1} and I_{A_2}, and have $Cov(I_{A_1}, I_{A_2}) = P(A_1A_2) - P(A_1)P(A_2) = \rho(I_{A_1}, I_{A_2})\sqrt{var(I_{A_1})var(I_{A_2})}$. Since $\rho(Z_1, Z_2) \geq 0$, we have $\rho(I_{A_1}, I_{A_2}) \geq 0$, and thus $P(A_1A_2) \geq P(A_1)P(A_2)$, which leads to $P(\max_{1\leq i\leq n} Z_i > t) \leq P(\max_{1\leq i\leq n} Z_i^* > t)$.

Next, we assume the conclusion holds for $n = k - 1$, and consider the case $n = k$. Define $M = \max_{1\leq i\leq k-1} Z_i$, $M^* = \max_{1\leq i\leq k-1} Z_i^*$, $A_M = \{M > t\}$ and $A_M^* = \{M^* > t\}$. We have $P(A_M) \leq P(A_M^*)$ by assumption.

Furthermore,

$$
\begin{aligned}
P(\max_{1\leq i\leq k} Z_i > t) &= P(\max\{M, Z_k\} > t) \\
&= P(A_M) + P(A_k) - P(A_MA_k) \\
&\leq P(A_M) + P(A_k) - P(A_M)P(A_k) \ \ \text{(By result for case } n = 2\text{)} \\
&= P(A_M)(1 - P(A_k)) + P(A_k) \\
&\leq P(A_M^*)(1 - P(A_k)) + P(A_k) \ \ \text{(By assumption)} \\
&= P(\max\{M^*, Z_k^*\} > t) = P(\max_{1\leq i\leq n} Z_i^* > t)
\end{aligned}
$$

Thus, the conclusion follows, and the proof is completed. □

Proof of Theorem 1

Proof. Now, we are ready to prove Theorem 1.

Part (1.a)

By Lemma S.1 and Assumption A, we have $T_1(x_i, y_j) \rightarrow_d \chi^2_1$ for each (x_i, y_j), as $n \rightarrow +\infty$, under H_0. Also, we may rewrite it as $T_1(x_i, y_j) \rightarrow_d Z_{ij}$, as $n \rightarrow +\infty$ with $Z_{ij} \sim \chi^2_1$. By the construction of the local statistic $T_1(x_i, y_j)$ that measures the local difference between two distributions, we have $\rho(T_1(x_i, y_j), T_1(x_k, y_l)) \geq 0$, for any (x_i, y_j) and (x_k, y_l).

Firstly, we construct random variable $T^*_1(x_i, y_j)$ such that (1) $T^*_1(x_i, y_j) =_d T_1(x_i, y_j)$; and (2) $T^*_1(x_i, y_j)$ for $i = 1, \cdots, n; j = 1, \cdots, m$ are independent. Let U_{ij} for $i = 1, \cdots, n; j = 1, \cdots, m$, i.i.d $\sim U(0, 1)$, and denote the distribution of $T_1(x_i, y_j)$ as $F_{ij}(t)$, then $T^*_1(x_i, y_j)$ can be constructed by $T^*_1(x_i, y_j) = F^{-1}_{ij}(U_{ij})$. Now, we have by Lemma S.4, for any $t > 0$,

$$P(\max_{1\leq i\leq n, 1\leq j\leq m} T_1(x_i, y_j) > t) \leq P(\max_{1\leq i\leq n, 1\leq j\leq m} T^*_1(x_i, y_j) > t) \tag{S4.1}$$

Next, since $T^*_1(x_i, y_j) =_d T_1(x_i, y_j) \rightarrow_d \chi^2_1$, as $n \rightarrow +\infty$, following Skorohod's representation theorem, there exits random variable $T^{**}_1(x_i, y_j)$ such that $T^{**}_1(x_i, y_j) \rightarrow_{a.s.} W_{ij}$ as $n \rightarrow +\infty$, $T^{**}_1(x_i, y_j) =_d T^*_1(x_i, y_j)$ and $W_{ij} \sim \chi^2_1$. Now, we have

$$P(\max_{1\leq i\leq n, 1\leq j\leq m} T^*_1(x_i, y_j) > t) = P(\max_{1\leq i\leq n, 1\leq j\leq m} T^{**}_1(x_i, y_j) > t) \tag{S4.2}$$

On the other hand, by construction of $T_1^{**}(x_i, y_j)$ and continuous mapping theorem, we have

$$\max_{1\leq i\leq n, 1\leq j\leq m} T_1^{**}(x_i, y_j) \rightarrow_{a.s.} \max_{1\leq i\leq n, 1\leq j\leq m} W_{ij} \tag{S4.3}$$

By Corollary S.1, we have $P(\max_{1\leq i\leq n, 1\leq j\leq m} W_{ij} > 8\log(nm)+1) \rightarrow 0$. Thus, we have $P(\max_{1\leq i\leq n, 1\leq j\leq m} T_1^{**}(x_i, y_j) > 8\log(nm)+1) \rightarrow 0$. Combining (S4.1) and (S4.2), we have $P\{\text{MAC}_1(X, Y) > 8\log(2nm)+1\} \rightarrow 0$.

Part (1.b)

By Lemma S.1, we have $T_2(\mathbf{x}_0, \mathbf{y}_0) \rightarrow \chi_3^2$, as $n, m \rightarrow +\infty$, under H_0. Similar to the proof of Part (1.a), we can show $\text{MAC}_2(\mathbf{X}, \mathbf{Y}) < 8\log(2nm)+3$ with probability going to 1.

Part (1.c)

By Lemma S.1, for any nonempty $s \subsetneqq S = \{1, \cdots, d\}$ we have $T_s(\mathbf{x}_0, \mathbf{y}_0) \rightarrow \chi_3^2$, as $n, m \rightarrow +\infty$, under H_0. Similar to the proof of Part (1.a), we have with probability going to 1, $\text{MAC}_3(\mathbf{X}, \mathbf{Y}) < 8\log(2^{d+1}nm)+3$.

Part (2)

Under H_1, let $p_i = P_i/n, q_i = Q_i/m, r_i = R_i/(n+m) = (P_i+Q_i)/(n+m)$, $(i = 1, 2)$, there is at least one point denoted by (x_0, y_0) such that $\Delta_{1i} = (p_i - r_i)^2 > 0$ and $\Delta_{2i} = (q_i - r_i)^2 > 0$ for $i = 1$ or $i = 2$. Without loss of generality, we assume it holds for $i = 1$, then we have $\text{MAC}_1(\mathbf{X}, \mathbf{Y}) > n\Delta_{11}/r_1 + m\Delta_{21}/r_1 \geq (n+m)\min(\Delta_{11}, \Delta_{21})/r_1 = c(n+m)$, with $c =$

$\min(\Delta_{11}, \Delta_{21})/r_1$.

Similarly, we can prove the corresponding results for $\text{MAC}_2(\mathbf{X}, \mathbf{Y})$ and $\text{MAC}_3(\mathbf{X}, \mathbf{Y})$.

□

S4.2 Proof of Corollary 1

Proof. Similar to the proof of part (a) in Theorem 1, we have with probability going to 1,

$$\max_{Z_j \in S_1^{*c}} \text{MAC}_1(Z_j^1, Z_j^0) < 8\log(2|S_1^{*c}|nm) + 1 < 8\log(2pnm) + 1.$$

If $p = e^{(n+m)^\eta}$ with $0 < \eta < 1$, we have $8\log(pnm) + 1 = O((n+m)^\eta)$, which is smaller than the lower bound of $\text{MAC}_1(Z_j^1, Z_j^0), Z_j \in S_1^*$. Thus, we have $P(M_1(\text{MAC}_1^1(n,m)) = S_1^*) \to 1$, as $n, m \to \infty$. □

S4.3 Proof of Corollary 2

Proof. The proof of (2.a) and (2.b) exactly follow the same manner as the proof of the first part of Corollary 1.

For convenience, we write $M_{21}(\text{MAC}_2^1(n,m))$ as M_{21}, $M_1(\text{MAC}_1^1(n,m))$ as M_1, and $M_{22}(\text{MAC}_2^2(n,m))$ as M_{22}.

$P(M_{21} = S_{21}^*) = P\{(M_{21} = S_{21}^*) \bigcap (M_1 = S_1^*)\} + P\{(M_{21} = S_{21}^*) \bigcap (M_1 \neq S_1^*)\}$.

By Corollary 1, we have $P(M_1 = S_1^*) \to 1$, as $n, m \to \infty$. Thus,

$$P(M_{21} = S^*_{21}) = P\{(M_{21} = S^*_{21}) \bigcap (M_1 = S^*_1)\} \to P(M^*_{21} = S^*_{21}) \to 1,$$

as $n, m \to \infty$.

where $M^*_{21} = \{X_{ij} = (Z_i, Z_j) : Z_i, Z_j \in S^{*c}_1$ and $\mathrm{MAC}_2(X^1_{ij}, X^0_{ij}) > \mathrm{MAC}^1_2(n, m)\}$.

Similarly, we can prove $P(M_{22} = S^*_{22}) \to 1$, as $n, m \to \infty$. □

Bibliography

Jianqing Fan. Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, 91(434):674–688, 1996.

Christopher Charles Heyde. *Selected Works of C.C. Heyde*. Springer Science & Business Media, 2010.

Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Robust multivariate nonparametric tests via projection-pursuit. *To appear in the Annals of Statistics, arXiv preprint arXiv:1803.00715*, 2020.

Qing Mai and Hui Zou. The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234, 2013.

Wen-Xin Zhou, Chao Zheng, and Zhen Zhang. Two-sample smooth tests for the equality of distributions. *Bernoulli*, 23(2):951–989, 2017.