Feature screening with kernel knockoffs

Benjamin Poignard
Osaka University/RIKEN AIP
bpoignard@econ.osaka-u.ac.jp

Peter Naylor
RIKEN AIP
peter.naylor@riken.jp

Héctor Climente-González
RIKEN AIP
hector.climente@riken.jp

Makoto Yamada
RIKEN AIP/Kyoto University
makoto.yamada@riken.jp

Abstract

This article analyses three feature screening procedures: Kendall's Tau and Spearman Rho (TR), Hilbert-Schmidt Independence Criterion (HSIC) and conditional Maximum Mean Discrepancy (cMMD), where the latter is a modified version of the standard MMD for categorical classification. These association measures are not based on any specific underlying model, such as the linear regression. We provide the conditions for which the sure independence screening (SIS) property is satisfied under a lower bound assumption on the minimum signal strength of the association measure. The SIS property for the HSIC and cMMD is established for given bounded and symmetric kernels. Within the high-dimensional setting, we propose a two-step approach to control the false discovery rate (FDR) using the knockoff filtering. The performances of the association measures are assessed through simulated and real data experiments and compared with existing competing screening methods.

1 INTRODUCTION

Feature selection, or variable selection, has gained much attention over the years and fostered the development of a rich literature, especially in predictive modelling. A standard approach to tackle the high-dimensionality issue is sparse modelling, which focuses on variable selection by discarding those that are unimportant for prediction. It aims to recover the true signal when the underlying model admits a sparse representation - this is the so-called "sparsity assumption" - by applying a penalty function to the objective function. A significant amount of literature has been dedicated to the devise of suitable penalty functions that can best fit the observed patterns and that satisfy particular properties: e.g., the large sample oracle property of Fan and Li (2001) for SCAD, of Zou (2006) for the adaptive LASSO or Poignard (2020) for Sparse Group LASSO; support recovery property under/without incoherence type conditions as in Loh and Wainwright (2017).

In light of the computational cost of these sparse methods and the issues regarding their statistical accuracy and algorithmic stability, Fan and Lv (2008) adopted the sure independent screening (SIS) viewpoint. This approach relies on marginal Pearson correlation learning and is designed for linear regressions with Gaussian predictors and responses so that it may not be robust to model misspecification. This gave rise to a broad range of studies dedicated to model-free SIS methods: the distance covariance of Székely et al. (2007) and Székely and Rizzo (2009), the distance correlation of Li et al. (2012b), the sup-HSIC of Balasubramanian et al. (2013), the Kolmogorov filter of Mai and Zou (2013, 2015), the projection correlation of Liu et al. (2020). As emphasized by Mai and Zou (2015), the purpose of variable selection is to identify the true sparse support, which may require intricate conditions such as incoherence/irrepresentable type ones, whereas feature screening aims to discover a majority of inactive features and it is thus less ambitious and requires weaker assumptions. More formally, feature screening attempts to find out the features that are independent with the target given the active features, where independence is assessed through a marginal utility.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

Our paper lies within the feature screening framework. We propose to investigate the SIS property for three association measures that do not rely on any underlying model assumption: the so-called TR measure, the conditional maximum mean discrepancy cMMD and the Hilbert–Schmidt Independence Criterion HSIC. The cMMD measure is used in the context of classification with categorical targets. Our contributions can be summarized as follows: we provide the conditions for which these association measures satisfy the SIS property; we propose a knockoff based filtering procedure to control for the false discovery rate (FDR); the proposed screening methods are a couple of orders of magnitude faster than that of existing state-of-the-art Liu et al. (2020) with holding comparable detection power. Our study shares a similar spirit with Barber and Candès (2019) and Liu et al. (2020), who devised knockoff procedures for FDR control within the high-dimensional setting $n < p$, with p the number of features and n the sample size. However, our work differs from these studies in two main aspects. First, we establish the SIS property for the new feature screening TR, which is particularly adapted to discrete outcomes, and the cMMD and HSIC for given bounded kernels. Second, we propose a two-step procedure for FDR control based on knockoff statistics which are adapted to the proposed association measures.

The rest of the article is organized as follows. In Section 2, we introduce the problem of feature screening. Section 3 provides the association measures to estimate the set of active features together with the conditions to satisfy the SIS property. In Section 4, we consider the knockoff-based FDR procedure. Finally, Section 5 contains numerical experiments for simulated and real data. All the intermediary results, proofs and additional figures are in the Supplementary Material.

2 FRAMEWORK

Throughout this paper, we consider a response variable Y and p features $(X_1, \cdots, X_p)$ among which we aim to discover the inactive ones through a marginal utility $\mathrm{D}(.,.)$ between Y and $X_k, 1 \leq k \leq p$. Our viewpoint is the sure independence screening one. In the same vein as in Li et al. (2012b), we denote by $\mathcal{S}$ the set of active features and by $\mathcal{I}$ the set of inactive features, respectively defined as

$$\begin{aligned}\mathcal{S} &:= \big\{k : \mathrm{F}(Y|X_1, \cdots, X_p) \\ &\quad \text{functionally depends on } X_k, 1 \leq k \leq p\big\}, \\ \mathcal{I} &:= \big\{k : \mathrm{F}(Y|X_1, \cdots, X_p) \\ &\quad \text{does not functionally depend on } X_k, 1 \leq k \leq p\big\},\end{aligned}$$

where $\mathrm{F}(Y|X_1, \cdots, X_p)$ is the probability distribution of $Y|X_1, \cdots, X_p$. We aim to find a majority of $\mathcal{I}$, that is given the set of active features $\{X_k, k \in \mathcal{S}\}$, we check that the features $\{X_k, k \in \mathcal{I}\}$ are independent of Y. To estimate $\mathcal{S}$ and screen out $\{X_k, k \in \mathcal{I}\}$, we consider the estimator of the marginal utility $\mathrm{D}(.,.)$, denoted by $\widehat{\mathrm{D}}(.,.)$, which quantifies the dependence between Y and covariate X_k. More formally, we compute $\widehat{\omega}_k = \widehat{\mathrm{D}}(\mathbf{Y}, \mathbf{X}_k)$ deduced from the sample of observations $\mathbf{Y} = (Y_1, \cdots, Y_n)$ and $\mathbf{X}_k = (X_{1k}, \cdots, X_{nk}), k = 1, \cdots, p$, and estimate $\mathcal{S}$ by the set

$$\widehat{\mathcal{S}}^{\lambda_n} := \big\{k : \widehat{\omega}_k \geq \lambda_n, \; k = 1, \cdots, p\big\}, \tag{1}$$

where λ_n is a threshold parameter that depends on the sample size and controls the number of selected active features. The convergence rate of λ_n is key to obtain the sure independence screening (SIS) property of the procedure, that is recovering with high probability the set $\mathcal{S}$ when estimating $\widehat{\mathcal{S}}$. In this paper, our contributions are threefold: (i) we prove the SIS property for three different measures $\mathrm{D}(.,.)$: $\mathrm{TR}(.,.), \mathrm{HSIC}(.,.), \mathrm{cMMD}(.,.)$; (ii) we establish the FDR control for these three measures through knockoff filtering in Section 4; (iii) our proposed estimators $\widehat{\omega}_k$ for feature screening provide a significant computational gain compared with existing methods.

3 SCREENING PROCEDURES

The key assumption to establish the sure screening property concerns the magnitude of the minimum $\mathrm{D}(Y, X_k), k \in \mathcal{S}$, and is specified as follows:

Assumption 1. *Let $0 < L_1 < \infty$ and $0 \leq \kappa < 1/2$, then $2L_1 n^{-\kappa} \leq \min_{k \in \mathcal{S}} \omega_k$ holds.*

This assumption is similar to condition (C2) of Li et al. (2012b) or Condition 1 (a) of Liu et al. (2020). It states that the association measure $\mathrm{D}(.,.)$ between Y and the X_k's, $k \in \mathcal{S}$, have a lower bound, whose rate is scaled by $\kappa < 1/2$. Hereafter, we denote by $k_0 = \mathrm{card}(\mathcal{S})$ the cardinality of $\mathcal{S}$.

3.1 Kendall's Tau and Spearman Rho (TR)

Spearman's ρ and Kendall's τ are extensively used to assess the existence of dependence between two continuous random variables. Li et al. (2012a) considered the Kendall's τ as a marginal utility for feature screening and established the SIS property. By extending the Kendall's τ measure to the discrete random variable case, Lu et al. (2018a) proposed a new measure, denoted as TR hereafter, which is a linear combination of the Spearman's ρ and Kendall's τ. They derived the large sample properties of the TR measure, whose sample version can be expressed as a U-statistic. We propose to use such a measure to perform variable screening when estimating the set $\mathcal{S}$. The TR measure between two random variables, say Y and X, is defined as

$$\text{TR}(Y,X) = 3\tau(Y,X) - 2\rho_s(Y,X),$$

where $\tau(Y,X)$ is the Kendall's τ and $\rho(Y,X)$ the Spearman's ρ of Y,X. Equipped with n samples $\mathbf{Y} = (Y_1,\cdots,Y_n)$, $\mathbf{X} = (X_1,\cdots,X_n)$, we denote by $\widehat{\tau}_u(\mathbf{Y},\mathbf{X})$ and $\widehat{\rho}_u(\mathbf{Y},\mathbf{X})$ the U-statistics counterparts of the Kendall's τ and the Spearman's ρ respectively, the U-statistics counterpart of the TR measure is

$$\begin{aligned}\widehat{\text{TR}}_u(\mathbf{Y},\mathbf{X}) &= 3\widehat{\tau}_u(\mathbf{Y},\mathbf{X}) - 2\widehat{\rho}_u(\mathbf{Y},\mathbf{X}) \\ &= 3\big(\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}^{n} \text{sgn}((X_i-X_j)(Y_i-Y_j))\big) \\ &\quad -2\big(\frac{12}{n(n-1)(n-2)}\sum_{\substack{i,j,l=1\\ i\neq j\neq k}}^{n} \mathbf{1}_{\big\{X_i>X_j,Y_i>Y_l\big\}} - 3\big).\end{aligned}$$

which can be written as a degree 3 symmetric kernel based U-statistic:

$$\widehat{\text{TR}}_u(\mathbf{Y},\mathbf{X}) = \frac{1}{n(n-1)(n-3)}\sum_{1\leq i<j<l\leq n}^{n} \varphi(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{Z}_l),$$

with $\boldsymbol{Z}_i = (Y_i,X_i)$ and $\varphi(.)$ the symmetric kernel of degree 3 given as

$$\begin{aligned}&\varphi(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{Z}_l) \\ &=2\big(\mathbf{1}_{\{X_i>X_j,Y_i>Y_j\}} + \mathbf{1}_{\{X_j>X_l,Y_j>Y_l\}} + \mathbf{1}_{\{X_l>X_i,Y_l>Y_i\}} \\ &\quad +\mathbf{1}_{\{X_j>X_i,Y_j>Y_i\}} + \mathbf{1}_{\{X_l>X_j,Y_l>Y_j\}} + \mathbf{1}_{\{X_i>X_l,Y_i>Y_j\}}\big) \\ &\quad -4\big(\mathbf{1}_{\{X_i>X_j,Y_i>Y_l\}} + \mathbf{1}_{\{X_j>X_l,Y_j>Y_i\}} + \mathbf{1}_{\{X_l>X_i,Y_l,Y_j\}} \\ &\quad +\mathbf{1}_{\{X_i>X_l,Y_i>Y_j\}}+\mathbf{1}_{\{X_j>X_i,Y_j>Y_l\}}+\mathbf{1}_{\{X_l>X_j,Y_l>Y_i\}}\big)+3.\end{aligned}$$

To deduce the values of the coefficients of the linear combination between $\widehat{\tau}_u(\mathbf{Y},\mathbf{X})$ and $\widehat{\rho}_u(\mathbf{Y},\mathbf{X})$, Lu et al. (2018a) considered the measure $\lambda_0\widehat{\tau}(X,Y) - (1-\lambda_0)\widehat{\rho}(X,Y)$, with $\lambda_0 = 3$ the value that minimizes the asymptotic variance of $\sqrt{n}(\lambda_0\widehat{\tau}(X,Y) - (1-\lambda_0)\widehat{\rho}(X,Y))$. Lu et al. (2018a) showed that under the null hypothesis $\mathbf{H0}: \mathbb{P}_{YX} = \mathbb{P}_Y\mathbb{P}_X$, then when $n\to\infty$, $\widehat{\text{TR}}_u(\mathbf{Y},\mathbf{X})$ converges to 0: this is the object of their Proposition 4.2, which is a particular case of $\widehat{\text{TR}}_u(\mathbf{Y},\mathbf{X})$ as the coefficients of the linear combination of the Spearman's and Kendall's measures are different from the TR(.,.) measure. To screen the active covariates $X_1,\cdots,X_p$ using the TR measure, we compute the set (1) with $\widehat{\omega}_k = |\widehat{\text{TR}}_u(\mathbf{Y},\mathbf{X}_k)|$; its population level counterpart is $\omega_k = |\text{TR}(Y,X_k)|$. The SIS property with TR is established in the next Theorem.

Theorem 3.1. *For any $\epsilon > 0$, there exists a finite constant $c_1 > 0$ such that*

$$\mathbb{P}\big(\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k| \geq \epsilon\big) \leq 2p\exp\big(-c_1 n\epsilon^2\big).$$

Under assumption 1, let $\lambda_n = L_2 n^{-\kappa} \leq \frac{1}{2}\min_{k\in\mathcal{S}}\omega_k$ with $L_2 > 0$, then there exists a finite $c'_1 > 0$ such that

$$\mathbb{P}\big(\mathcal{S}\subseteq\widehat{\mathcal{S}}^{L_2 n^{-\kappa}}\big) \geq 1 - 2k_0\exp\big(-c'_1 n^{1-2\kappa}\big).$$

3.2 conditional Maximum Mean Discrepancy (cMMD)

We now consider an association measure that allows to assess whether the conditional distribution $X_k|Y, 1\leq k\leq p$ is equal to the distribution X_k in the context of a categorical Y. To do so, we rely on the setting developed by Ke and Yin (2020), which specifies such a measure via the Maximum Mean Discrepancy (MMD). First, let us briefly provide the framework to define the MMD distance, that will also be useful when defining HSIC. Let $\mathcal{X}$ be a metric space and $\mathcal{H}$ a Hilbert space of functions $f:\mathcal{X}\to\mathbb{R}$. $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) induced by the inner product $\langle .,.\rangle$ if there exists a function $\phi:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ such that $\forall\boldsymbol{x}\in\mathcal{X}$, $\phi(\boldsymbol{x},.)\in\mathcal{H}$ and $\forall f\in\mathcal{H},\forall\boldsymbol{x}\in\mathcal{X},\langle f,\phi(\boldsymbol{x},.)\rangle = f(\boldsymbol{x})$. For any probability measure $\mathbb{P}$ defined on $\mathcal{X}$, the mean $\mu(\mathbb{P})\in\mathcal{H}$ is defined as $\mathbb{E}[f(X)] = \langle f,\mu(\mathbb{P})\rangle$ for any $f\in\mathcal{H}$, where the random variable X is sampled from $\mathcal{X}$. Equipped with the RKHS, we consider two random variables $Y\sim\mathbb{P}_Y$ and $X\sim\mathbb{P}_X$ that take values on $(\mathcal{Y},\mathcal{B}_y)$ and $(\mathcal{X},\mathcal{B}_x)$, respectively, where $\mathcal{Y},\mathcal{X}$ are two separable metrics, and $\mathcal{B}_y,\mathcal{B}_x$ are Borel σ-algebras. Then, $(\mathcal{Y}\times\mathcal{X},\mathcal{B}_y\times\mathcal{B}_x)$ is measurable, and the joint distribution is defined as $\mathbb{P}_{Y\,X}$, which assigns values to the product space $(\mathcal{Y}\times\mathcal{X},\mathcal{B}_y\times\mathcal{B}_x)$. We define the symmetric and bounded kernel $\phi(.,.)$. The MMD corresponds to the distance in $\mathcal{H}$ of the means, or equivalently is the distance between two probability measures: more details on RKHS and MMD can be found in Gretton et al. (2012). Now by Lemma 6 of Gretton et al. (2012), the population level MMD association measure is given as

$$\begin{aligned}\text{MMD}^2(Y,X_k) &= \mathbb{E}_{YY'}[\varphi(Y,Y')] \\ &\quad+\mathbb{E}_{X_kX'_k}[\varphi(X_k,X'_k)] - 2\mathbb{E}_{YX_k}[\varphi(Y,X_k)].\end{aligned}$$

Now rather than comparing the distributions X_k and Y, Ke and Yin (2020) consider the distance between the distributions $X_k|Y$ and X_k, that we call hereafter cMMD - "c" for conditional -, defined as

$$\begin{aligned}&\text{cMMD}^2(Y,X_k) := \mathbb{E}_Y[\gamma^2(Y)] \\ &\quad= \mathbb{E}_Y\big[\mathbb{E}_{X_kX'_k}[\varphi(X_k,X'_k)|Y,Y']\big] - \mathbb{E}_{X_kX'_k}[\varphi(X_k,X'_k)],\end{aligned}$$

where $\mathbb{E}_Y[.]$ is the expectation with respect to Y, $\mathbb{E}_{X_kX'_k}[.|Y=y,Y'=y]$ denotes the conditional expectation of $X_k,X'_k|Y=y,Y'=y$, with (Y',X'_k) an independent copy of (Y,X_k) and $\gamma^2(Y)$ is defined as

$$\begin{aligned}\gamma^2(Y) &= \mathbb{E}_{X_kX'_k}[\varphi(X_k,X'_k)] + \mathbb{E}_{X_k}\big[\mathbb{E}_{X'_k}[\varphi(X_k,X'_k)|Y']\,|Y\big] \\ &\quad - 2\mathbb{E}_{X_k}\big[\mathbb{E}_{X'_k}[\varphi(X_k,X'_k)]\,|Y\big].\end{aligned}$$

The variable $\gamma^2(Y)$ is defined in the same spirit as the MMD distance, but rather than comparing the equality of two distributions, $\text{cMMD}^2(Y, X_k)$ allows to assess the independence between Y and X_k: indeed, by Theorem 4 of Ke and Yin (2020), when $\omega_k := \text{cMMD}^2(Y, X_k) = 0$, then X_k and Y are independent. Now equipped with n samples of (Y_i, X_i), to estimate (1) in the context of a categorical Y, we assume that Y has L levels such that $Y = l, 1 \leq l \leq L$ with probability $\pi_l \in [0,1], 1 \leq l \leq L$ and each level has n_l observations. Then we compute for all $1 \leq k \leq p$ the statistic $\widehat{\omega}_k = \widehat{\text{cMMD}}^2_v(\mathbf{Y}, \mathbf{X}_k)$, which is specified as

$$\widehat{\omega}_k = \sum_{l=1}^{L} \widehat{\pi}_l \frac{1}{n_l^2} \sum_{i,j \in \mathcal{E}_l} \varphi(X_{ik}, X_{jk}) - \frac{1}{n^2} \sum_{i,j=1}^{n} \varphi(X_{ik}, X_{jk}).$$

with $\mathcal{E}_l = \{i : Y_i = l\}$, n_l the number of observations for the l-th level and $\widehat{\pi}_l = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i = l}$. We now establish the SIS property for the $\text{cMMD}^2(.,.)$.

Theorem 3.2. *Assume $\|\varphi\|_\infty = c$, then for any $0 < \epsilon < 1$ and $n \geq Mc/\epsilon$ with $M > 0$ a finite constant, the following bound holds:*

$$\begin{aligned}\mathbb{P}\big(\max_{1 \leq k \leq p} |\widehat{\omega}_k - \omega_k| \geq \epsilon\big) \\ \leq \quad 2p\big(\sum_{l=1}^{L} \exp\big(-c_1 n_l \epsilon^2\big) + \exp\big(-c_2 n \epsilon^2\big)\big),\end{aligned}$$

with $c_1, c_2 > 0$ finite constants (independent of n). Under assumption 1, let $\lambda_n = L_2 n^{-\kappa} \leq \frac{1}{2}\min_{k \in \mathcal{S}} \omega_k$ with $L_2 > 0$, then there exists constants $c_1', c_2' > 0$ such that

$$\begin{aligned}&\mathbb{P}\big(\mathcal{S} \subseteq \widehat{\mathcal{S}}^{L_2 n^{-\kappa}}\big) \\ &\geq 1 - 2k_0\big(L \exp\big(-c_1' n^{1-2/\kappa}\big) + \exp\big(-c_2' n^{1-2/\kappa}\big)\big).\end{aligned}$$

Several comments can be emphasized:

(i) For the sake of simplification, we proved the SIS property under the bounded kernel assumption. The latter condition can be relaxed in favor of moment conditions such as sub-exponential tail conditions. In that case, a slower convergence rate for the bound over $\max_{1 \leq k \leq p} |\widehat{\omega}_k - \omega_k|$ would be obtained.

(ii) The quantity L is assumed fixed. In the same spirit as in Lu et al. (2021), who considered a different version of cMMD in the context of grouped categorical variables and for the distance kernel only, we can consider a diverging number of classes $L = O(n^\mu), 0 \leq \mu < 1/2$.

(iii) The cMMD distance can accommodate continuous variables. In that case, its estimator is a degree 5 U-statistic closely related to the HSIC measure: see subsection 3.2. of Ke and Yin (2020). Another interesting association measure close to MMD and for the general case (Y, X) continuous is the kernel partial correlation (KPC) measure developed by Huang et al. (2020) defined as $\omega_k = \mathbb{E}[\text{MMD}^2(\mathbb{P}_Y^{|X_k}, \mathbb{P}_Y)]/\mathbb{E}[\text{MMD}^2(\delta_Y, \mathbb{P}_Y)]$, where $\mathbb{P}_Y^{|X_k}$ denotes the conditional distribution of $Y|X_k$, $\mathbb{P}_Y$ the (un)conditional distribution of Y, δ_Y the Dirac measure of Y. Here, ω_k is a particular case of the KPC measure as highlighted in their remark 2.4. The sample estimator proposed by Huang et al. (2020) is $\widehat{\omega}_k = \text{tr}\big(M^\top \widetilde{\mathbf{K}}_{\mathbf{Y}} M\big)/\text{tr}\big(\widetilde{\mathbf{K}}_{\mathbf{Y}}\big)$, where $M = \widetilde{\mathbf{K}}_{\mathbf{X}_k}(\widetilde{\mathbf{K}}_{\mathbf{X}_k} + n\epsilon I_n)^{-1}$ with $\epsilon > 0$ a fixed tuning parameter, $\widetilde{\mathbf{K}}_{\mathbf{Y}} = \mathbf{H}\mathbf{K}_{\mathbf{Y}}\mathbf{H}$, $\widetilde{\mathbf{K}}_{\mathbf{X}_k} = \mathbf{H}\mathbf{K}_{\mathbf{X}_k}\mathbf{H}$, where $\mathbf{K}_{\mathbf{Y}} = (\varphi(Y_i, Y_j))_{1 \leq i,j \leq n}$, $\mathbf{K}_{\mathbf{X}_k} = (\varphi(X_{ik}, X_{jk}))_{1 \leq i,j \leq n}$ and $\mathbf{H} = I_n - \frac{1}{n}\iota\iota^\top$. If we are in a position to derive an exponential bound on both the numerator and denominator for a given bounded kernel, then the SIS property would follow. We leave this as a future research.

3.3 Hilbert-Schmidt Independence Criterion (HSIC)

The last stone of our three-part study on feature screening to estimate (1) is the Hilbert-Schmidt Independence Criterion (HSIC) measure. A significant literature has been dedicated to HSIC: see, e.g., Gretton et al. (2005); Song et al. (2012) for a theoretical analysis of the HSIC measure as a feature selection method; Poignard and Yamada (2020) for the use HSIC in the context of penalised HSIC based mRMR. The HSIC measure is a covariance measure in RKHS, which is defined in Subsection 3.2. We define two symmetric bounded kernels $\phi(.,.), \psi(.,.)$ on the spaces $\mathcal{Y}$ and $\mathcal{X}$. The HSIC measure of $\mathbb{P}_{YX_k}$ is then given as

$$\begin{aligned}\text{HSIC}(Y, X_k) = \mathbb{E}_{YY'X_kX_k'}[\phi(Y,Y')\psi(X_k,X_k')] \\ +\mathbb{E}_{YY'}[\phi(Y,Y')]\mathbb{E}_{X_kX_k'}[\psi(X_k,X_k')] \\ -2\mathbb{E}_{YX_k}[\mathbb{E}_{Y'}[\phi(Y,Y')]\mathbb{E}_{X_k'}[\psi(X_k,X_k')]],\end{aligned}$$

where (Y', X') is an i.i.d. copy of (Y, X), and $\mathbb{E}_{XX'}[.]$ (resp. $\mathbb{E}_X[.]$) is the expectation defined over X, X' (resp. X). Then $\omega_k = \text{HSIC}(Y, X_k)$ and we compute (1) as $\widehat{\omega}_k = \widehat{\text{HSIC}}_v(\mathbf{Y}, \mathbf{X}_k)$, which is the V-statistic based estimator of ω_k defined as

$$\begin{aligned}\widehat{\text{HSIC}}_v(\mathbf{Y}, \mathbf{X}_k) = \frac{1}{n^2}\sum_{i,j=1}^{n} L_{ij}K_{ij} \\ +\frac{1}{n^4}\sum_{i,j,m,l=1}^{n} L_{ij}K_{ml} - \frac{2}{n^3}\sum_{i,j,m=1}^{n} L_{ij}K_{im},\end{aligned}$$

with $L_{ij} = \phi(Y_i, Y_j)$, $K_{ij} = \psi(X_{ik}, X_{jk})$. As highlighted by Gretton et al. (2005) in their Theorem 1, the bias of such a V-statistic is of order $O(n^{-1})$. We now provide the conditions for which SIS holds for given bounded and symmetric kernels $\phi(.,.), \psi(.,.)$.

Theorem 3.3. *Assume $\|\phi\|_\infty = c_1, \|\psi\|_\infty = c_2$, $c_1, c_2 < \infty$, then for $0 < \epsilon < 1$ and $n \geq L\eta/\epsilon$ with $L > 0$ finite, $\eta = c_1 c_2$, for c a finite constant:*

$$\mathbb{P}(\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k| \geq \epsilon) \leq p\, O(\exp\big(-cn\epsilon^2\big)).$$

Under assumption 1, let $\lambda_n = L_2 n^{-\kappa} \leq \frac{1}{2}\min_{k\in\mathcal{S}} \omega_k$ with $L_2 > 0$, then there exists a finite constant c' such that

$$\mathbb{P}\big(\mathcal{S} \subseteq \widehat{\mathcal{S}}^{L_2 n^{-\kappa}}\big) \geq 1 - k_0\, O(\exp\big(-c' n^{1-2\kappa}\big)).$$

It is worth mentioning the following remarks:

(i) Balasubramanian et al. (2013) proved the SIS property for the sup-HSIC measure, where the key technical step is the derivation of an exponential bound over sup-HSIC using the McDiarmid's inequality and a symmetrization argument, implying a different exponential bound for $\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k|$ compared to ours.

(ii) As in the cMMD case, we assumed bounded kernels. Such an assumption can be relaxed in favor of moment conditions.

(iii) As an alternative, we will consider in our applications the normalised version of the HSIC measure $\widehat{\omega}_k = \widehat{\text{HSIC}}_v(\mathbf{Y}, \mathbf{X}_k)/\big(\widehat{\text{HSIC}}_v(\mathbf{Y}, \mathbf{Y})\widehat{\text{HSIC}}_v(\mathbf{X}_k, \mathbf{X}_k)\big)^{1/2}$. Equipped with the bound on $\widehat{\text{HSIC}}_v(\mathbf{Y}, \mathbf{X}_k)$ according to Theorem 3.3, similar bounds on $\widehat{\text{HSIC}}_v(\mathbf{Y}, \mathbf{Y}), \widehat{\text{HSIC}}_v(\mathbf{X}_k, \mathbf{X}_k)$ can be straightforwardly deduced so that by Lemma S.2 of Liu et al. (2020), one can obtain an exponential bound on $\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k|$ and deduce the SIS property.

4 OPTIMAL SCREENING AND FDR

Now we propose to bound the false discovery rate (FDR) while performing variable screening following the knockoff+ method developed by Barber and Candès (2015). Since the seminal work of Barber and Candès (2015) on the knockoff procedure and its applications for FDR control, a broad range of studies has been flourishing on the extensions and applications of the knockoff method: Candès et al. (2018) devised the Model-X knockoff in the context of a random design matrix; Barber and Candès (2019) addressed the issue of knockoff-based FDR control within the high-dimensional setting and emphasized how knockoffs can be applied to non-sparse signals; Romano et al. (2020) devised a Model-X knockoff framework almost model-free with applications to unsupervised datasets; in the same vein as Barber and Candès (2019), Fan et al. (2020) and Liu et al. (2020) considered two-step approaches for FDR control; Lu et al. (2018b) applied the knockoff filtering to deep neural neural networks. The details on the construction of the knockoff variables $\widetilde{X}_j, j = 1, \cdots, p$ can be found in Section A of the Supplementary Material.

The knockoff+ method for feature selection can be broken down into the following three steps:

(i) Construct the knockoff variables $\widetilde{X}_j$ for all $j = 1, \cdots, p$. The knockoff variable is an artificial version of the original variable. To do so, two methods can be performed: the equicorrelated method or the semi-definite program method.

(ii) For each pair of original variable X_j and knockoff variable $\widetilde{X}_j$, compute a statistic W_j such that a large value of W_j gives evidence that j is a true signal. The definition of this statistic depends on the association measure. More precisely, for any $j = 1, \cdots, p$, $W_j := W_j^{\text{tr}}$ (resp. $W_j^{\text{cmmd}}, W_j^{\text{hsic}}$), which is defined according to the association measures for which we established the SIS property:

$$\begin{aligned} W_j^{\text{tr}} &= |\text{TR}(Y, X_j)| - |\text{TR}(Y, \widetilde{X}_j)|, \\ W_j^{\text{cmmd}} &= \text{cMMD}^2(Y, X_j) - \text{cMMD}^2(Y, \widetilde{X}_j), \\ W_j^{\text{hsic}} &= \text{HSIC}(Y, X_j) - \text{HSIC}(Y, \widetilde{X}_j). \end{aligned}$$

The statistic W_j serves as a signal on how important the original variable is compared to its knockoff version. For such a signal to work, W_j must satisfy the sufficiency and the anti-symmetry properties. Replacing each association measure by its empirical counterpart, we consider the estimator $\widehat{W}_j$. A higher value of $\widehat{W}_j$ gives evidence that the distribution of Y depends on X_j. Should X_j be inactive, then $|\widehat{W}_j|$ is close to zero.

(iii) Define a selection rule to carry out feature selection. To do so, we specify a data-dependent threshold $T(\alpha)$ similar to equation (13) of Barber and Candès (2019) as

$$T(\alpha) = \min\big\{t \in \mathcal{W} :\ \frac{1 + \text{card}(\widehat{W}_j \leq -t)}{\text{card}(\widehat{W}_j \geq t) \vee 1} \leq \alpha\big\}, \tag{2}$$

and $T(\alpha) = +\infty$ should this set be empty and where $\mathcal{W} = \big\{|\widehat{W}_j| :\ 1 \leq j \leq p\big\} \setminus \{0\}$ is the set of unique nonzero values reached by the $|\widehat{W}_j|$'s. Then the active set is defined as

$$\widehat{\mathcal{S}}(\alpha) = \big\{1 \leq j \leq p :\ \widehat{W}_j \geq T(\alpha)\big\}.$$

The way the statistic $\widehat{W}_j$ is built determines the success of the procedure controlling the FDR and should be carefully specified according to the association measure. Its specification should satisfy the so-called flip-coin property or anti-symmetry property: swapping the pair $(\mathbf{X}_j, \widetilde{\mathbf{X}}_j)$ only changes the sign of $\widehat{W}_j$ and keeps the sign of $\widehat{W}_{j'}, j' \neq j$ unchanged. More precisely, the signs of the $\widehat{W}_j, j \in \mathcal{S}^c$ are i.i.d. random variables and independent from $|\widehat{W}_j|$ for any $j = 1, \cdots, p$ and from $\operatorname{sgn}(W_j)$ for $j \in \mathcal{S}$. Such a property is formalized in the following Lemma, which is in the same vein as in Lemma 1 of Barber and Candès (2015) or Lemma 1 of Liu et al. (2020).

Lemma 4.1. *Let $\epsilon_j \in \{\pm 1\}, j = 1, \cdots, p$, be a sign sequence such that $\epsilon_j \perp \widehat{W}_j$ for any j. Each ϵ_j satisfies $\forall j \in \mathcal{S}, \epsilon_j = 1$ and $\forall j \in \mathcal{S}^c, \epsilon \overset{iid}{\sim} \pm 1$. Then $(\widehat{W}_1, \cdots, \widehat{W}_p) \overset{d}{=} (\epsilon_1 \widehat{W}_1, \cdots, \epsilon_p \widehat{W}_p)$.*

This Lemma shows that, given $(|\widehat{W}_1|, \cdots, |\widehat{W}_p|)$, then

$$\operatorname{card}\big(j \in \mathcal{S}^c : \ \widehat{W}_j \geq t\big) \ \overset{d}{=} \ \operatorname{card}\big(j \in \mathcal{S}^c : \ \widehat{W}_j \leq -t\big),$$

$t > 0$ and both follow the same Binomial distribution.

Finally, at threshold t, the false discovery proportion (FDP) is estimated as the ratio

$$\widehat{\text{FDP}}(t) = \frac{\operatorname{card}\big(j : \ \widehat{W}_j \leq -t\big)}{\operatorname{card}\big(j : \ \widehat{W}_j \geq t\big) \vee 1}.$$

One key hurdle in step (i) is the sample size requirement $2p < n$, which is often not satisfied. We will hence rely on a standard data splitting approach in the same vein as in Section 4 of Barber and Candès (2019) so that the feature select ion procedure can be performed in two steps: first we carry out feature selection on a sub-sample; then, using only the selected variables, we perform the knockoff procedure on the remaining samples. Our proposed screening method can be summarized as follows:

(i) Split the sample into two parts n_0 and n_1 such that $n_0 + n_1 = n$. Then define the vectors of observations $\mathbf{X}^{(0)} \in \mathbb{R}^{n_0 \times p}, \mathbf{Y}^{(0)} \in \mathbb{R}^{n_0}$ and $\mathbf{X}^{(1)} \in \mathbb{R}^{n_1 \times p}, \mathbf{Y}^{(1)} \in \mathbb{R}^{n_1}$ so that $\mathbf{Y} = (\mathbf{Y}^{(0)\top}, \mathbf{Y}^{(1)\top})^\top$ and $\mathbf{X} = (\mathbf{X}^{(0)\top}, \mathbf{X}^{(1)\top})^\top$.

(ii) Perform feature selection by ranking in descending order the features on the sub-sample n_0 using the three association measures. Then select the s_0 features such that $2s_0 < n_1$. The set of selected features is then denoted as $\widehat{\mathcal{S}}_0$, whose cardinality is s_0.

(iii) Construct the knockoff $\widetilde{\mathbf{X}}^{(1)}_{\widehat{\mathcal{S}}_0}$ from $\mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0}$, where $\mathbf{X}^{(1)} = (\mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0}, \mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0^c})$. Then for each $j = 1, \cdots, s_0$, let $\mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0,j}$ (resp. $\widetilde{\mathbf{X}}^{(1)}_{\widehat{\mathcal{S}}_0,j}$) the j-th column of $\mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0}$ (resp. $\widetilde{\mathbf{X}}^{(1)}_{\widehat{\mathcal{S}}_0}$), we compute the statistic for each pair of original variable and knockoff variable

$$\begin{aligned}
\widehat{W}_j^{\text{tr}} &= |\widehat{\text{TR}}_u(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0,j})| - |\widehat{\text{TR}}_u(\mathbf{Y}^{(1)}, \widetilde{\mathbf{X}}^{(1)}_{\widehat{\mathcal{S}}_0,j})|,\\
\widehat{W}_j^{\text{cmmd}} &= \widehat{\text{cMMD}}_v(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0,j}) - \widehat{\text{cMMD}}_v(\mathbf{Y}^{(1)}, \widetilde{\mathbf{X}}^{(1)}_{\widehat{\mathcal{S}}_0,j}),\\
\widehat{W}_j^{\text{hsic}} &= \widehat{\text{HSIC}}_v(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0,j}) - \widehat{\text{HSIC}}_v(\mathbf{Y}^{(1)}, \widetilde{\mathbf{X}}^{(1)}_{\widehat{\mathcal{S}}_0,j}).
\end{aligned}$$

Then for a fixed α, we use (2) to estimate the set of active features

$$\widehat{\mathcal{S}}(\alpha) = \big\{1 \leq j \leq s_0 : \ \widehat{W}_j \geq T(\alpha)\big\}.$$

We define the sure screening event as

$$\mathcal{E} = \big\{\mathcal{S} \subseteq \widehat{\mathcal{S}}_0, \ 2\operatorname{card}\big(\widehat{\mathcal{S}}_0\big) < n_1\big\}.$$

Depending on ω_k, the probability of $\mathcal{E}$ admits different lower bounds: see Theorems 3.1, 3.2, 3.3. The next result establishes that given the selection event $\mathcal{E}$, the proposed screening procedures with knockoff based features control for the FDR for a given level α.

Theorem 4.2. *Let $\widehat{W} = (\widehat{W}_1, \cdots, \widehat{W}_p)$ satisfy the anti-symmetry and sufficiency property. For any level $0 < \alpha < 1$, let the data-dependent threshold $T(\alpha) = \min\big\{t \in \mathcal{W} : \ \frac{1 + \operatorname{card}(\widehat{W}_j \leq -t)}{\operatorname{card}(\widehat{W}_j \geq t) \vee 1} \leq \alpha\big\}$, with $\mathcal{W} = \big\{|\widehat{W}_j| : \ 1 \leq j \leq p\big\} \setminus \{0\}$ and the set of active features $\widehat{\mathcal{S}}(\alpha) = \big\{1 \leq j \leq p : \ \widehat{W}_j \geq T(\alpha)\big\}$. Then*

$$\mathbb{E}\Big[\frac{\operatorname{card}(j : \ j \in \mathcal{S}^c \ \text{ and } \ j \in \widehat{\mathcal{S}}(\alpha)}{\operatorname{card}(j : \ j \in \widehat{\mathcal{S}}(\alpha)) \vee 1} | \mathcal{E}\Big] \leq \alpha.$$

Theorem 4.2 is based on a two-step procedure: first, screening on $\mathbf{Y}^{(0)}, \mathbf{X}^{(0)}$ for a given measure, whose success is ensured in Section 3; construction of the knockoff features from $\mathbf{X}^{(1)}$ and assessment of FDR using the statistics $\widehat{W}_j^{\text{tr}}, \widehat{W}_j^{\text{cmmd}}, \widehat{W}_j^{\text{hsic}}$. This two-step approach is in the same vein as in Barber and Candès (2019), Liu et al. (2020) or Fan et al. (2020).

5 APPLICATIONS

5.1 Simulated experiments

We compare the finite sample screening performances of $\text{TR}, \text{cMMD}, \text{HSIC}$ with the Distance Correlation (DC) of Li et al. (2012b), the Projection Correlation (PC) of Liu et al. (2020) and Pearson coefficient. Based on several data generating processes (DGPs), we generate Y from linear and non-linear transformation of some active covariates $X_k, k \in \mathcal{S}$ with $\mathcal{S}$ known. We use these screening methods to rank in descending order the selected covariates and consider the minimum model size containing all the active features. This procedure is replicated 200 times so that we obtain a minimum model size averaged over these 200 batches. The minimum model size, say $\mathcal{M}$, corresponds to the minimum model size to include all active features. The screening performance is measured through the quantiles $(5\%, 25\%, 50\%, 75\%, 95\%)$ of $\mathcal{M}$.

5.1.1 Setup

We denote by $X \in \mathbb{R}^{n\times p}$ the matrix of covariates containing $n \in \mathbb{N}^*$ samples and $p \in \mathbb{N}^*$ features and by $Y \in \mathbb{R}^n$ the target output. Each input sample is drawn from a gaussian distribution, $X_i \sim \mathcal{N}(0_p, \Sigma)$ where $\Sigma = (\sigma_{i,j})_{p\times p}$ and $\sigma_{i,j} = c^{|i-j|}$, where we set $c = 0.5$, except stated otherwise. We denote by ε the error term, which is set by default as $\varepsilon \sim \mathcal{N}_{\mathbb{R}}(0,1)$ except stated otherwise. We denote by $\mathcal{T}(k)$ the Student distribution with k degrees of freedom. We denote by $\mathbf{1}_n$ the n-dimensional vector of one's. Equipped with these notations, we consider the following DGPs:

Linear models:

$$\begin{aligned}
&\textbf{1.a}: \; Y = \beta_0.X_{\mathcal{S}} + \varepsilon, \text{ where } \beta_0 = (1,2,4,8).\\
&\textbf{1.b}: \; Y = \mathbf{1}_{10}.X_{\mathcal{S}} + \varepsilon.\\
&\textbf{1.c}: \; Y = \mathbf{1}_{10}.X_{\mathcal{S}} + \varepsilon, \text{ where } \varepsilon \sim \mathcal{T}(2).
\end{aligned}$$

Non linear models:

$$\begin{aligned}
\textbf{2.a}: \; Y = \;& 5X_1 + 2\sin(\pi X_2/2) + 2X_3\mathbb{1}\{X_3 > 0\} + \\
& 2\exp\{5X_4\} + \varepsilon.\\
\textbf{2.b}: \; Y = \;& 3X_1 + 3X_2^3 + 3X_3^{-1} + 5\mathbb{1}\{X_4 > 0\} + \varepsilon.\\
\textbf{2.c}: \; Y \sim \;& \mathcal{P}(\lambda), \text{ where } \mathcal{P} \text{ is the Poisson distribution}\\
& \text{and } \lambda = \exp\{\mathbf{1}_{10}.X_{\mathcal{S}}\}.
\end{aligned}$$

Categorical data:

$$\begin{aligned}
\textbf{3.a}: \; Y = \;& \mathbb{1}\{\text{logit}(\mathbf{1}_{10}.X_{\mathcal{S}}) > 0.5\},\\
& \text{where } \text{logit}(x) = (1 + \exp\{-x\})^{-1}.\\
\textbf{3.b}: \; Y = \;& \text{Same as } \textbf{3.a} \text{ except that } c = 0.\\
\textbf{3.c}: \; Y = \;& \begin{cases} 0 \text{ if } Y < 0,\\ \lceil 0.5Y \rceil \text{ if } Y^* \in [0;8] \text{ and } Y^* = \mathbf{1}_{10}.X_{\mathcal{S}} + \varepsilon,\\ 5 \text{ if } 8 \leq Y^*. \end{cases}
\end{aligned}$$

For each DGP we vary the sample size $n \in \{100, 500, 1000\}$ and the number of features $p \in \{100, 500, 5000\}$, except for $n = 1000$ paired with $p = 100$ and $p = 500$. When running the knockoff procedure, we set $n_0 = \lfloor 0.3n \rfloor$, s_0 is set respectively to n and takes the following values $\{50, 300, 100\}$.

Finally, both cMMD and HSIC require the specification of a symmetric and bounded kernel. For a random variable Z, we considered the following kernels:

$$\begin{aligned}
\textbf{Linear:}\; \varphi(Z_i, Z_j) = \;& Z_i.Z_j,\\
\textbf{Gaussian:}\; \varphi(Z_i, Z_j) = \;& \exp\left(-\frac{|Z_i - Z_j|^2}{2 * \sigma_z}\right),\\
\textbf{Distance:}\; \varphi(Z_i, Z_j) = \;& |Z_i|^\gamma + |Z_j|^\gamma - |Z_i - Z_j|^\gamma,
\end{aligned}$$

where γ is set to 1 in our experiments. For the Gaussian kernel, σ_z is the widths of the kernel defined according to the median heuristic (Sriperumbudur et al., 2009), i.e. $\sigma_z = 2^{-1/2}\text{median}(\{|Z_i - Z_j|^2\}_{i,j=1}^n)$.

5.1.2 Screening performances

For each association measure, we check its ability and efficiency to retrieve the true active features. In particular, for the 200 batches, we record the smallest integer m such that the true k_0 features are contained within the top m scores given by the screening. For DGPs **1.a**, **2.a** and **2.b** we have $k_0 = 4$ whereas for the rest we have $k_0 = 10$.

Figure 1 displays the screening results for DGP **3.a**: as expected, all association measure screen well and notably Pearson, TR and the linear and distance kernels have a smaller standard error. In addition, Table 1 shows the high average computation time of PC. Due to this fact, we were not able to perform experiments with PC on large datasets such as the case $(n, p) = (1000, 5000)$. In Figure 1 in the Supplementary Material, we display the time and memory usage for each measure with respect to its input size. More precisely, the complexity of TR is bounded by Kendall's τ which is in $O(n \log n)$ whereas the complexity of PC is bounded by $O(n^3)$.

n	DC	PC	HSIC	cMMD	TR
100	0.45	48	0.47	2.3	0.77
200	0.71	451	1.3	4.5	0.79
300	0.97	1527	2.4	7.2	0.78

Table 1: Average processing time (milliseconds) of one sample of size $(n, 1)$ over 500 runs. Vectors generated randomly.

We reported in the Supplementary Material, Section D, from Figure 2 to Figure 10, the screening performances for the other DGPs. The minimum model size with a random selection procedure for $p = 100$ features is averaged to 95, for $p = 500$ to 475 and for $p = 5000$ to 4750, we do not show this distribution on the plots to not unnecessarily spread out the y-axis to prevent squashing of the other distributions. For all

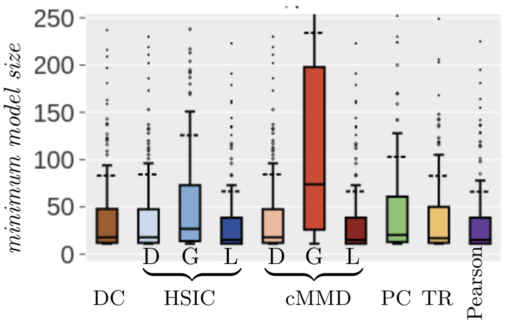

Figure 1: Minimum model size, with $n = 500$ and $p = 5000$ on DGP **3.a**. Kernels set as L: linear kernel; G: Gaussian; D: Distance.

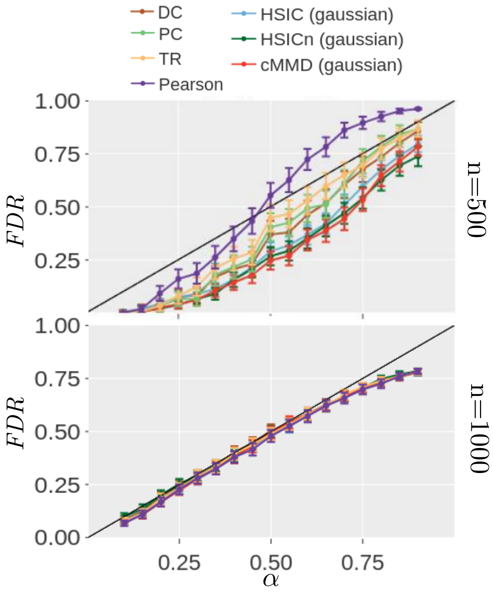

Figure 2: FDR control with $n = 500$ or $n = 1000$, and $p = 5000$ on DGP **3.a**. HSICn: normalised HSIC.

metrics and datasets, we first observe that they perform better than a random selection by a far margin. The performances of DC and PC are in line with Li et al. (2012b) and Liu et al. (2020).

For the linear based DGP, whose results are depicted in Figures 2, 3 and 4, as expected, most association measures perform well and nearly pinpoint the correct covariates when n is larger than 500. TR seems to be the least effective for these models. Figures 5, 6 and 7 concern the non-linear based DGP. As the objective is more difficult, the minimum model size naturally increases. PC and TR outperform the other measures on **2.a** and **2.c**. The performances for the case of categorical outputs are reported in Figures 8, 9 and 10: the best screening results are achieved by PC, TR and HSIC, cMMD with the linear kernel.

5.1.3 FDR control

We now assess whether the knockoff procedure controls for the FDR. The $\widehat{W}_j$ statistic for DC is set the same way as Liu et al. (2020) for PC. $\widehat{W}_j$ is set as the difference of the absolute values for the Pearson coefficient. In Figure 2, we show the FDR with respect to α for DGP **3.a** with $(n, p) = (500, 5000)$. The FDR rate only goes above α for the Pearson measure when the number of samples is lower and the knockoff procedure seems well in control for the others. In the Supplementary Material, Figures 11 to 16, we confirm that this observation also holds for the other DGPs, except for HSIC and DC for the non-linear DGPs when $n = 500$ and $p = 5000$ in Figures 14, 15 and 16. This excess is negated when n increases to 1000. As for Figures 20 to 25, we vary the kernel and check that the FDR control still holds for HSIC and cMMD. We observe that only the linear kernel exceeds the alpha value for the non-linear DGP and for the last two categorical DGPs.

Importantly, the FDR control should be put in perspective with the number of empty sets that the procedure returns. We recall that a model returns a perfect FDR score of 0 when the model keeps no features. Figure 29 displays the percentage of models returning an empty set. We computed this statistic by pooling the results from all the DGPs. It would seem that the empirical probability to select an empty set is proportional and decreases with respect to α. For the lowest α, the probability can be as high as 80%.

5.1.4 Hyper parameter choices

In light of the experiments we previously performed, we provide some guidelines on how to choose the following hyper parameters: the association measure $\mathrm{D}(.,.)$; the kernels φ, or $\phi(.,.), \psi(.,.)$, if relevant; n_0 the number of samples used for screening; and s_0 the number of features retained after screening.

The choice of the relevant measure/kernel is data dependent and targets the relationship between the active covariates $X_k, k \in \mathcal{S}$ and Y. If the underlying relationship is linear, a linear kernel is better suited. If the relationship is non-linear, a distance kernel or the TR measure should be a closer fit. However, in real world applications, the relationship is usually unknown and, ideally, the choice of $\mathrm{D}(.,.)$ should be motivated by experiments on withheld data. To do so, we released a Python Package that implements all measures and kernels considered and where custom kernels and measures can easily be integrated for better fits.

In the Supplementary Material, in Figures 30 and 31, we show the minimum model size, FDR control and

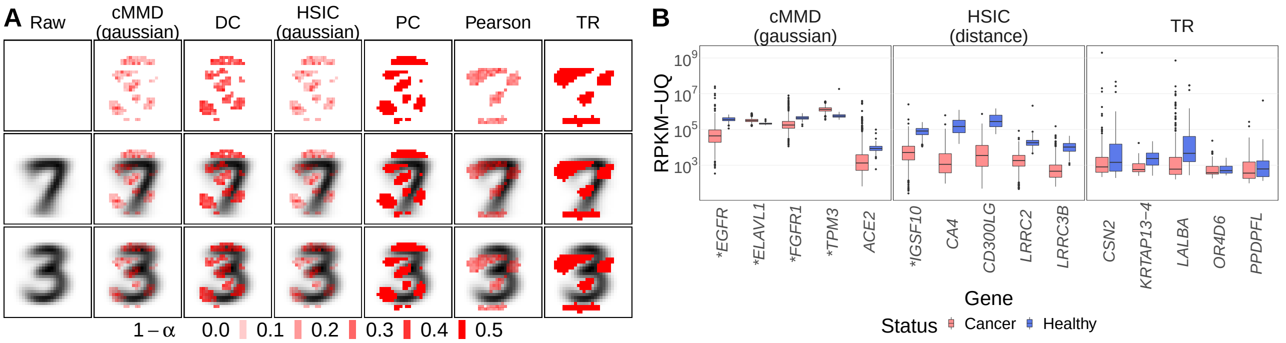

Figure 3: **(A)** Pixels selected on 8 train-test splits of MNIST (red). The shade of red indicates the lowest α threshold at which the pixels were selected (stronger means lower). The top row shows the pixels on a white background, while the two lower rows overlay exactly the same image to the average three, and the average seven. **(B)** Gene expression of the 5 genes with highest $\widehat{W}_j$ values among those selected at $\alpha = 0.1$ using different association measures. The 5 genes marked with an asterisk are labeled as oncogenes, tumor suppressors or candidate cancer drivers in the NCG 7.0 database Dressler et al. (2021).

percentage of returned empty sets for increasing values of n_0. In particular, for all values of n_0, the FDR remains under-control. Furthermore, on one hand, we notice that higher values of n_0 are correlated with a better screening. Naturally, a larger n_0 induces more samples for the screening step. On the other hand, higher values of n_0 are correlated to a higher percentage of the empty sets being returned. Indeed a smaller n_1 results in a $\widehat{W}_j$ statistic being based on fewer samples and will therefore be less stable. We recommend setting n_0 to anything between 40% and 60% of n.

Finally, we recommend setting s_0 to the highest value possible with respect to n_1 to increase the chance that the true features pass the screening step. But one could choose to shrink it for computational reasons.

5.2 Real-world applications

We applied the proposed procedure to the MNIST dataset (LeCun et al., 2010), selecting pixels that are good at discriminating sevens from threes while keeping the FDR under control. We included 500 images of threes, and 500 images of sevens, each of them consisting of 784 pixels. For each value of α and association measure, we ran our pipeline on a different training set containing 87.5% of the samples. For each of the association measures, we screened the best 100 features on 10% of the training samples. A visual examination shows that all association measures select reasonable discriminant pixels between the two classes: see Figure 3-(A). This was supported by the high classification testing accuracies of random forest classifiers trained on the selected pixels, as highlighted in Figure 32.

Additionally, we used our proposed procedure to discover relevant genes in breast carcinoma. To do so, we worked on the BRCA cohort from The Cancer Genome Atlas, whose data was provided by the TCGA Research Network. We considered RNA-seq measures of gene expression (RPKM-UQ) of 56 602 genes obtained from 1 102 samples from primary tumors and 113 samples from normal breast tissue. We searched for good predictors exclusively among the 18 868 protein coding genes using the proposed protocol and different association measures. The expression of the respective best 5 genes is available in Figure 3-(B). As we show, 5 of those 15 genes have been linked to cancer, supporting the notion that the protocol is selecting relevant features. As detailed in the Supplementary Material, Section E, this trend holds when we look at all 261 selected genes. This is also supported by the high classification test accuracy of a random forest (see Figure 33), trained in an identical setup as for MNIST, and using all 56 602 genes.

Code availability

In addition to the theoretical and methodological developments, for the sake of reproducible research, we make the code for all experiments publicly available in the following Github repository `https://github.com/PeterJackNaylor/knockoff-MMD-HSIC`.

Acknowledgments

We thank all the reviewers for their constructive and insightful comments that significantly improved this paper. B.P. was supported by Start-Support KAKENHI 19K23193 and MEXT KAKENHI 20H04243. H.C.-G. was supported by the RIKEN Special Postdoctoral Researcher Program. M.Y. was supported by MEXT KAKENHI 20H04243 and partly supported by MEXT KAKENHI 21H04874. We are also thankful for the RAIDEN computing system and its support team at the RIKEN AIP, which we used for our experiments.

References

K. Balasubramanian, B. Sriperumbudur, and G. Lebanon. Ultrahigh dimensional feature screening via RKHS embeddings. In *AISTATS, Proceedings of Machine Learning Research*, volume 31, pages 126–134, 2013.

L. D. Barber and E.J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

L. D. Barber and E.J. Candès. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, Layton. R., J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

E.J. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-x' knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B*, 80(3):551–577, 2018.

P. Di Tommaso, M. Chatzou, E.W. Floden, P.P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

Lisa Dressler, Michele Bortolomeazzi, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo, Lucia Montorsi, Neshika Wijewardhane, Dimitra Repana, Joel Nulsen, Jacki Goldman, Marc Pollit, Patrick Davis, Amy Strange, Karen Ambrose, and Francesca D. Ciccarelli. Comparative assessment of genes driving cancer and somatic evolution. preprint, Cancer Biology, August 2021. URL `http://biorxiv.org/lookup/doi/10.1101/2021.08.31.458177`.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Y. Fan, E. Demirkaya, G. Li, and J. Lv. Rank: Large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 115 (529):362–379, 2020.

A. Gretton, O. Bousquet, Alex. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005.

A. Gretton, K.M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13: 723–773, 2012.

Z. Huang, N. Deb, and B. Sen. Kernel partial correlation coefficient — a measure of conditional dependence. *arxiv*, (2012.14804), 2020.

C. Ke and X. Yin. Expected conditional characteristic function-based measures for testing independence. *Journal of the American Statistical Association*, 115 (530):985–996, 2020.

G.M Kurtzer, V. Sochat, and M.W. Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017.

Y. LeCun, C. Cortes, and C.J. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877, 2012a.

R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012b.

W. Liu, Y. Ke, Liu J., and R. Li. Model-free feature screening and fdr control with knockoff features. *Journal of the American Statistical Association*, 2020.

P.-L. Loh and M.J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

D. Lu, L. Zhang, X. Wang, and L. Song. Some new measures of dependence for random variables based on spearman's ρ and kendall's τ. *Journal of Nonparametric Statistics*, 30(4):860–883, 2018a.

J. Lu, L. Lin, and W. Wang. Partition-based feature screening for categorical data via rkhs embeddings. *Computational Statistics and Data Analysis*, 157, 2021.

Y.Y. Lu, J. Lv, Y. Fan, and W.S. Noble. Deeppink: Reproducible feature selection in deep neural networks. *arXiv*, (1809.01185), 2018b.

Q. Mai and H. Zou. The kolmogorov filter for variable screening in high dimensional binary classification. *Biometrika*, 100(1):229–234, 2013.

Q. Mai and H. Zou. The fused kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497, 2015.

B. Poignard. Asymptotic theory of the adaptive sparse group lasso. *Annals of the Institute of Statistical Mathematics*, 72:297–328, 2020.

B. Poignard and M. Yamada. Sparse hilbert-schmidt independence criterion regression. In *AISTATS, Proceedings of Machine Learning Research*, volume 108, pages 538–548, 2020.

Y. Romano, M. Sesia, and E.J. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

R.J. Serfling. *Approximation theorems of mathematical statistics.* New York: Wiley, 1980.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13: 1393–1434, 2012.

B.K. Sriperumbudur, K. Fukumizu, A. Gretton, G.R. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS*, 2009.

G. J. Székely and M.L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4): 1236–1265, 2009.

G. J. Székely, M.L. Rizzo, and N.K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 (476):1418–1429, 2006.

Supplementary Material:
Feature screening with kernel knockoffs

This Supplementary Material is organized as follows. Section A provides a short review on the knockoff construction. All the proofs are included in Section B. Section C details the implementation procedures we relied on. Section D displays the figures for the screening performances, FDR control for all DGPs but DGP **3.a**, and additional figures related to the real data experiments. Finally, Section E provides further comments on the breast cancer data.

A KNOCKOFF TOOLKIT

We propose a brief review of the knockoff filter following the work of Barber and Candès (2015); Candès et al. (2018). Let $\mathbf{X} = (X_1, \cdots, X_p)$ the p-dimensional vector of covariates. The new vector of variables $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \cdots, \widetilde{X}_p)$ is a knockoff of $\mathbf{X}$ if the following two properties are satisfied:

(i) for any $\mathcal{S} \subseteq \{1, \cdots, p\}$, $(\mathbf{X}, \widetilde{\mathbf{X}})_{\text{swap}(\mathcal{S})} \overset{d}{=} (\mathbf{X}, \widetilde{\mathbf{X}})$, where $(\mathbf{X}, \widetilde{\mathbf{X}})_{\text{swap}(\mathcal{S})}$ is obtained from $(\mathbf{X}, \widetilde{\mathbf{X}})$ when swapping the entries X_j and $\widetilde{X}_j$ for each $j \in \mathcal{S}$. This is the so-called pairwise exchangeable property.

(ii) the distribution of $\widetilde{\mathbf{X}}$ is independent of $Y|\mathbf{X}$.

When those two properties are satisfied, then $\widetilde{\mathbf{X}}$ is a model-X knockoff for $\mathbf{X}$. Under the assumption of Gaussian covariates, constructing such model-X knockoffs can be straightforwardly carried out following subsection 3.1.1 of Candès et al. (2018). However, the distribution of $\mathbf{X}$ may be unknown and the condition $(\mathbf{X}, \widetilde{\mathbf{X}})_{\text{swap}(\mathcal{S})} \overset{d}{=} (\mathbf{X}, \widetilde{\mathbf{X}})$ is challenging to satisfy. To circumvent this issue, Candès et al. (2018) propose an approximate construction for the knockoff $\widetilde{\mathbf{X}}$ in their subsection 3.4: their so-called second order model-X knockoff require moment conditions for $(\mathbf{X}, \widetilde{\mathbf{X}})_{\text{swap}(\mathcal{S})}$ and $(\mathbf{X}, \widetilde{\mathbf{X}})$. More precisely, the expectations of $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ should match. Let $\text{Var}(\mathbf{X}) = \Sigma$, then the second order condition - that is equality of the covariances - is equivalent to

$$\text{Var}((\mathbf{X}, \widetilde{\mathbf{X}})) = \mathbf{G},\ \mathbf{G} = \begin{pmatrix} \Sigma & \Sigma - \text{diag}(\mathbf{s}) \\ \Sigma - \text{diag}(\mathbf{s}) & \Sigma \end{pmatrix},$$

where $\mathbf{s} = (s_k, k = 1, \cdots, p) \in \mathbb{R}^p$ is a p-dimensional vector such that $\mathbf{G} \succeq 0$, that is positive semidefinite. Here $\text{diag}(\mathbf{s})$ denotes a $p \times p$ diagonal matrix with diagonal components $s_1, \cdots, s_p$. To obtain a suitable $\mathbf{s}$, in the same spirit as Barber and Candès (2015), Candès et al. (2018) propose two methods. The first approach is the *equicorrelated* construction, which specifies $\mathbf{s}$ as

$$\forall k = 1, \cdots, p,\ s_k = 2\lambda_{\min}(\Sigma) \wedge 1,$$

where $\lambda_{\min}(A)$ is the minimum eigenvalue of a square symmetric matrix A. The second approach is the *semidefinite program*, where the suitable $\mathbf{s}$, say $\overline{\mathbf{s}}$, satisfies the convex program:

$$\min_{\mathbf{s}} \sum_{k=1}^{p} |1 - s_k|,\ \text{ subject to } \forall k, s_k \geq 0,\ \Sigma - \text{diag}(\mathbf{s}) \succeq 0,$$

where $A \succeq 0$ meaning A being semidefinite positive.

The equicorrelated and semidefinite program methods are not directly applicable for a large p, respectively for the following reasons: $\lambda_{\min}(\Sigma)$ is close to zero when p becomes large so that the equicorrelated approach has low power; when p is large, the convex semidefinite program is computationally expensive. To fix these issues, Candès et al. (2018) propose a two-step based procedure for the semidefinite program, the so-called approximate semidefinite program. In the first step, one considers:

$$\min_{\mathbf{s}} \sum_{k=1}^{p} |1-s_k|, \;\; \text{subject to} \;\; \forall k, s_k \geq 0, \; \Sigma^* - \text{diag}(\mathbf{s}) \succeq 0,$$

for a suitable Σ^*, for example Σ^* can be calibrated as an m-block-diagonal approximation of Σ. Denoting by $\bar{\mathbf{s}}^*$ the optimal solution of the latter problem, in the second step, given such a $\bar{\mathbf{s}}*$, one considers:

$$\max_{\gamma} \gamma, \;\; \text{subject to} \;\; 2\Sigma - \text{diag}(\gamma\bar{\mathbf{s}}*),$$

a problem satisfied by $\overline{\gamma}$. Then one sets the approximate semidefinite program based $\mathbf{s}$ as $\bar{\mathbf{s}} = \overline{\gamma}\,\bar{\mathbf{s}}^*$.

In our two-step procedure described in Section 4, where we first screen the features on the subsample $\mathbf{Y}^{(0)}, \mathbf{X}^{(0)}$ so that we are left with s_0 features such that $2s_0 < n_1$, where $s_0 = \text{card}(\widehat{\mathcal{S}}_0)$ is the cardinality of the set containing the active features estimated over the subsample n_0. As a consequence, we are in a position to apply both the equicorrelated or semidefinite program on $\mathbf{X}^{(1)}_{\widehat{\mathcal{S}}_0} \in \mathbb{R}^{n_1 \times s_0}$.

B Proofs

The SIS property we establish rely on the derivation of exponential bounds on $\max_{1\leq k\leq p} \; |\widehat{\omega}_k - \omega_k|$. To do so, we rely on the exponential inequality for U-statistics of Theorem 5.6.1.A of Serfling (1980).

Proof of Theorem 3.1

Proof. First, let us focus on the exponential bound

$$\forall \epsilon > 0, \; \mathbb{P}\big(\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k| \geq \epsilon\big) \leq 2d \exp\big(-c_1 n \epsilon^2\big).$$

In view of the degree 3 symmetric kernel $\varphi(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_l)$, which is a combination of indicator functions and each of them being bounded by 1, we thus deduce that $-9 \leq \varphi(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_l) \leq 9$. Hence, by Theorem 5.6.1.A of Serfling (1980), and using the symmetry of the U-statistic, we obtain

$$\forall \epsilon > 0, \forall 1 \leq k \leq p, \;\; \mathbb{P}\big(|\widehat{\mathrm{TR}}_u(\mathbf{Y}, \mathbf{X}_k) - \mathrm{TR}(Y, X_k)| \geq \epsilon\big) \leq 2\exp\big(-\frac{2}{81}\lfloor n/3 \rfloor \epsilon^2\big).$$

We then conclude by union bound

$$\mathbb{P}\big(\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k| \geq \epsilon\big) \leq 2p\exp\big(-c_1 n \epsilon^2\big), \tag{3}$$

with $c_1 > 0$ a finite constant.

To establish the screening property, if $\mathcal{S} \nsubseteq \widehat{\mathcal{S}}^{\lambda_n}$, then there exists $k \in \mathcal{S}$ such that $\widehat{\omega}_k < L_2 n^{-\kappa}$. By the minimum signal assumption $2L_1 n^{-\kappa} \leq \min\limits_{k\in\mathcal{S}} \omega_k$, then $|\widehat{\omega}_k - \omega_k| > L_1 n^{-\kappa}$ for $k \in \mathcal{S}$. Hence $\big\{\mathcal{S} \nsubseteq \widehat{\mathcal{S}}^{\lambda_n}\big\} \subseteq \big\{|\widehat{\omega}_k - \omega_k| > L_1 n^{-\kappa}\big\}$ for $k \in \mathcal{S}$. Let us denote $\mathcal{A}_n = \big\{\max\limits_{1\leq k\leq d} |\widehat{\omega}_k - \omega_k| \leq L_1 n^{-\kappa}\big\}$: we have $\mathcal{A}_n \subset \big\{\mathcal{S} \subseteq \widehat{\mathcal{S}}^{\lambda_n}\big\}$. On this set for any $k \in \mathcal{S}$ and under the minimum signal condition,

$$|\widehat{\omega}_k| \geq |\omega_k| - |\widehat{\omega}_k - \omega_k| \geq L_1 n^{-\kappa}.$$

As a consequence, by inequality (3), we deduce $\mathbb{P}(\mathcal{A}_n^c) \leq 2k_0 \exp(-c_1' n^{1-2\kappa})$ with $c_1' > 0$ finite. Finally, taking $\lambda_n \leq L_1 n^{-\kappa} \leq \frac{1}{2}\min\limits_{k\in\mathcal{S}} \omega_k$,

$$\mathbb{P}(\mathcal{S} \subseteq \widehat{\mathcal{S}}^{\lambda_n}) \geq \mathbb{P}(\mathcal{A}_n) = 1 - \mathbb{P}(\mathcal{A}_n^c) \geq 1 - 2k_0 \exp(-c_1' n^{1-2\kappa}).$$

□

Proof of Theorem 3.2

Proof. The empirical counterpart of $\omega_k = \text{cMMD}(Y, X_k)$, denoted by $\widehat{\text{cMMD}}(\mathbf{Y}, \mathbf{X}_k), k = 1, \cdots, p$, is given as:

$$\widehat{\omega}_k = \sum_{l=1}^{L} \widehat{\pi}_l \frac{1}{n_l^2} \sum_{i,j \in \mathcal{E}_l} \varphi(X_{ik}, X_{jk}) - \frac{1}{n^2} \sum_{i,j=1}^{n} \varphi(X_{ik}, X_{jk}) := T_{1n}(\mathbf{Y}, \mathbf{X}_k) + T_{2n}(\mathbf{X}_k),$$

with $\mathcal{E}_l = \{i : Y_i = l\}$, n_l the number of observations for the l-th level and $\widehat{\pi}_l = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i = l}$ is the empirical probability counterpart. Now $T_{1n}(\mathbf{Y}, \mathbf{X}_k)$ can be expressed as $T_{1n}(\mathbf{Y}, \mathbf{X}_k) := \sum_{l=1}^{L} \widehat{\pi}_l T_{1n,l}(\mathbf{Y}, \mathbf{X}_k)$, with obvious notations. As for the population level counterpart, we have

$$\mathbb{E}_Y[\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')|Y, Y']] = \sum_{l=1}^{L} \pi_l T_{1,l}, \ T_{1,l} = \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')|Y = l, Y' = l].$$

The objective is to bound both $T_{1n}(\mathbf{Y}, \mathbf{X}_k)$ and $T_{2n}(\mathbf{X}_k)$ in $\widehat{\omega}_k$. First, we consider $T_{1n}(\mathbf{Y}, \mathbf{X}_k)$, which can be bounded in a similar fashion as Lu et al. (2021), who considered a grouping structure for the empirical estimator of cMMD. We have

$$\begin{aligned} &|T_{1n}(\mathbf{Y}, \mathbf{X}_k) - \mathbb{E}_Y[\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')|Y, Y']]| \\ \leq\ & \sum_{l=1}^{L} \widehat{\pi}_l |T_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| + |\sum_{l=1}^{L} (\widehat{\pi}_l - \pi_l) T_{1,l}| \leq \max_{1 \leq l \leq L} |T_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| + |\sum_{l=1}^{L} (\widehat{\pi}_l - \pi_l) T_{1,l}|. \end{aligned}$$

Now we have the relationship between the V-statistic and U-statistic of $T_{1n,l}(\mathbf{Y}, \mathbf{X}_k)$ given as:

$$n_l^2 T_{1n,l}(\mathbf{Y}, \mathbf{X}_k) = n_l(n_l - 1) U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) + n_l U_{0n,l}(\mathbf{Y}, \mathbf{X}_k), \text{ or } T_{1n,l}(\mathbf{Y}, \mathbf{X}_k) = \frac{n_l - 1}{n_l} U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) + \frac{1}{n_l} U_{0n,l}(\mathbf{Y}, \mathbf{X}_k),$$

with $U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) = \frac{1}{n_l(n_l - 1)} \sum_{i \neq j \in \mathcal{E}_l} \varphi(X_{ik}, X_{jk})$ and $U_{0n,l}(\mathbf{Y}, \mathbf{X}_k) = \frac{1}{n_l} \sum_{i \in \mathcal{E}_l} \varphi(X_{ik}, X_{ik})$. Then by the bound assumption on φ, for any $l \in L$, for any $0 < \epsilon < 1$, for n sufficiently large such that $n_l \geq Kc/\epsilon$ and $T_{1,l}/n_l \leq \epsilon$, $U_{0n,l}(\mathbf{Y}, \mathbf{X}_k)/n_l \leq \epsilon$, then

$$\begin{aligned} &\mathbb{P}(|T_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| \geq 3\epsilon) \\ =\ & \mathbb{P}(|\frac{n_l - 1}{n_l} U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) + \frac{1}{n_l} U_{0n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| \geq 3\epsilon) = \mathbb{P}(|\frac{n_l - 1}{n_l}(U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}) - \frac{1}{n_l} T_{1,l}| \geq 2\epsilon) \\ \leq\ & \mathbb{P}(\frac{n_l - 1}{n_l} |U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| \geq 2\epsilon - |\frac{1}{n_l} T_{1,l}|) \leq \mathbb{P}(|U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| \geq \epsilon). \end{aligned}$$

By an application of Theorem 5.6.1.A of Serfling (1980), since $U_{1n,l}(\mathbf{Y}, \mathbf{X}_k)$ is a degree 2 kernel, we deduce

$$\mathbb{P}(U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l} \geq \epsilon) \leq \exp(-2\lfloor n_l/2 \rfloor \epsilon^2/c^2).$$

The bound holds for deviations in the opposite direction: hence

$$\forall 0 < \epsilon < 1, \ \mathbb{P}(|U_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| \geq \epsilon) \leq 2\exp(-2\lfloor n_l/2 \rfloor \epsilon^2/c^2).$$

By union bound, we conclude

$$\forall 0 < \epsilon < 1, \ \mathbb{P}(\max_{1 \leq l \leq L} |T_{1n,l}(\mathbf{Y}, \mathbf{X}_k) - T_{1,l}| \geq \epsilon) \leq 2\sum_{l=1}^{L} \exp(-2\lfloor n_l/2 \rfloor \epsilon^2/9c^2).$$

Moreover, for $0 < \epsilon < 1$,

$$\mathbb{P}(|\sum_{l=1}^{L} (\widehat{\pi}_l - \pi_l) T_{1,l}| \geq \epsilon) = \mathbb{P}(|\sum_{l=1}^{L} (\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_i = l}) T_{1,l} - \pi_l T_{1,l}| \geq \epsilon) \leq 2\exp(-2n\epsilon^2),$$

by an application of Theorem 5.6.1.A of Serfling (1980). We then deduce for any $0 < \epsilon < 1$ that

$$\begin{aligned}
&\mathbb{P}(|T_{1n}(\mathbf{Y},\mathbf{X}_k) - \mathbb{E}_Y[\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')|Y, Y']]| \geq \epsilon) \\
\leq \quad & \mathbb{P}(\max_{1\leq l\leq L}|T_{1n,l}(\mathbf{Y},\mathbf{X}_k) - T_{1,l}| \geq \epsilon/2) + \mathbb{P}(|\sum_{l=1}^{L}(\widehat{\pi}_l - \pi_l)T_{1,l}| \geq \epsilon/2) \\
\leq \quad & 2\sum_{l=1}^{L}\exp(-\lfloor n_l/2\rfloor\epsilon^2/18c^2) + 2\exp(-n\epsilon^2/2).
\end{aligned}$$

We now focus on bounding $T_{2n}(\mathbf{X}_k)$, which is a standard V-statistic. Let us denote $\widehat{U}_{2n}(\mathbf{X}_k) = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\varphi(X_{ik}, X_{jk})$, so that

$$n^2 T_{2n}(\mathbf{X}_k) = n(n-1)U_{2n}(\mathbf{X}_k) + nU_{0n}, \;\text{ or }\; T_{2n}(\mathbf{X}_k) = (n-1)/nU_{2n}(\mathbf{X}_k) + U_{0n}/n,$$

with $U_{0n} = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_{ik}, X_{ik})$. Then for $0 < \epsilon < 1$, take n such that $n \geq Mc/\epsilon$, then $U_{0n}/n \leq \epsilon$ and $\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]/n \leq \epsilon$. We deduce

$$\begin{aligned}
&\mathbb{P}(|T_{2n}(\mathbf{X}_k) - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq 3\epsilon) \\
= \quad & \mathbb{P}(|\frac{n-1}{n}U_{2n}(\mathbf{X}_k) + \frac{1}{n}U_{0n} - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq 3\epsilon) \\
\leq \quad & \mathbb{P}(|\frac{n-1}{n}U_{2n}(\mathbf{X}_k) - \frac{n-1}{n}\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')] - \frac{1}{n}\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq 2\epsilon) \\
\leq \quad & \mathbb{P}(|\frac{n-1}{n}\big(U_{2n}(\mathbf{X}_k) - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]\big) \geq 2\epsilon - |\frac{1}{n}\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]|) \\
\leq \quad & \mathbb{P}(|U_{2n}(\mathbf{X}_k) - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq \epsilon).
\end{aligned}$$

Then by Theorem 5.6.1.A of Serfling (1980), and using the symmetry of the U-statistics, we deduce

$$\mathbb{P}(|T_{2n}(\mathbf{X}_k) - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq 3\epsilon) \leq \mathbb{P}(|U_{2n}(\mathbf{X}_k) - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq \epsilon) \leq 2\exp\big(-2\lfloor n/2\rfloor\epsilon^2/c^2\big).$$

Putting the pieces together,

$$\begin{aligned}
&\mathbb{P}(|\widehat{\text{cMMD}}_v^2(\mathbf{Y},\mathbf{X}_k) - \text{cMMD}^2(Y, X_k)| \geq \epsilon) \\
= \quad & \mathbb{P}(|T_{1n}(\mathbf{Y},\mathbf{X}_k) + T_{2n}(\mathbf{X}_k) - \mathbb{E}_Y\big[\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')|Y, Y']\big] - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq \epsilon) \\
\leq \quad & \mathbb{P}(|T_{1n}(\mathbf{Y},\mathbf{X}_k) - \mathbb{E}_Y\big[\mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')|Y, Y']\big]| \geq \epsilon/2) + \mathbb{P}(|T_{2n}(\mathbf{X}_k) - \mathbb{E}_{X_k X_k'}[\varphi(X_k, X_k')]| \geq \epsilon/2) \\
\leq \quad & 2\sum_{l=1}^{L}\exp(-\lfloor n_l/2\rfloor\epsilon^2/72c^2) + 2\exp(-n\epsilon^2/8) + 2\exp\big(-\lfloor n/2\rfloor\epsilon^2/2c^2\big).
\end{aligned}$$

Using the exponential bound on $\widehat{\text{cMMD}}_v^2(\mathbf{Y},\mathbf{X}_k)$, the bound over $\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k|$ is straightforward. To establish the sure screening property, we follow the same steps as in Theorem 3.1. □

Proof of Theorem 3.3

Proof. The V-statistic and U-statistic estimators of $\text{HSIC}(Y, X_k), k = 1, \cdots, p$, are respectively defined as

$$\begin{aligned}
&\widehat{\text{HSIC}}_v(\mathbf{Y},\mathbf{X}_k) \\
= \quad & \frac{1}{n^2}\sum_{i,j=1}^{n}L_{ij}K_{ij} + \frac{1}{n^4}\sum_{i,j,m,l=1}^{n}L_{ij}K_{ml} - \frac{2}{n^3}\sum_{i,j,m=1}^{n}L_{ij}K_{im} := \widehat{S}_1^v + \widehat{S}_2^v - 2\widehat{S}_3^v,
\end{aligned}$$

$$\begin{aligned}
&\widehat{\mathrm{HSIC}}_u(\mathbf{Y},\mathbf{X}_k)\\
&= \frac{1}{n(n-1)}\sum_{(i,j)\in\mathbf{i}_2^n} L_{ij}K_{ij} + \frac{1}{n(n-1)(n-2)(n-3)}\sum_{(i,j,m,l)\in\mathbf{i}_4^n} L_{ij}K_{ml} - \frac{2}{n(n-1)(n-2)}\sum_{(i,j,m)\in\mathbf{i}_3^n} L_{ij}K_{im}\\
&:= \widehat{S}_1^u + \widehat{S}_2^u - 2\widehat{S}_3^u,
\end{aligned}$$

where $L_{ij} = \phi(Y_i, Y_j)$, $K_{ij} = \psi(X_{ik}, X_{jk})$, which are bounded kernels by assumption and $\mathbf{i}_m^n$ is the set all m-tuples drawn without replacement from the set $\{1,\cdots,n\}$. First, we have the relationship:

$$\widehat{S}_1^v - \widehat{S}_1^u = \frac{1}{n}\widehat{S}_0^u - \frac{1}{n}\widehat{S}_1^u \Leftrightarrow \widehat{S}_1^v = \frac{n-1}{n}\widehat{S}_1^u + \frac{1}{n}\widehat{S}_0^u,\text{ where } \widehat{S}_0^u = \frac{1}{n}\sum_{i=1}^n L_{ii}K_{ii}.$$

Let us now focus on $\widehat{S}_3^v$. We have the relationship

$$\widehat{S}_3^v - \widehat{S}_3^u = -\frac{3}{n}\widehat{S}_3^u + \frac{1}{n^3}\sum_{(i,j)\in\mathbf{i}_2^n}\big(K_{ii}L_{ij} + K_{ij}L_{ii} + K_{ij}L_{ij}\big) + O(n^{-2}),$$

so that

$$\widehat{S}_3^v = \frac{n-3}{n}\widehat{S}_3^u + \frac{n-1}{n^2}\widehat{T}^u + O(n^{-2}),\text{ with } \widehat{T}^u := \frac{1}{n(n-1)}\sum_{(i,j)\in\mathbf{i}_2^n}\big(K_{ii}L_{ij} + K_{ij}L_{ii} + K_{ij}L_{ij}\big).$$

Finally, we have

$$\widehat{S}_2^v - \widehat{S}_2^u = -\frac{6}{n}\widehat{S}_2^u + \frac{1}{n^4}\sum_{(i,j,m)\in\mathbf{i}_3^n}\big(K_{ii}L_{jm} + 4K_{ij}L_{im} + K_{ij}L_{mm}\big) + O(n^{-2}).$$

We then deduce

$$\widehat{S}_2^v = \frac{n-6}{n}\widehat{S}_2^u + \frac{(n-1)(n-2)}{n^3}\widehat{P}^u + O(n^{-2}),\text{ with } \widehat{P}^u := \frac{1}{n(n-1)(n-2)}\sum_{(i,j,m)\in\mathbf{i}_3^n}\big(K_{ii}L_{jm} + 4K_{ij}L_{im} + K_{ij}L_{mm}\big).$$

First, let us bound $\widehat{S}_1^v$. By the bound assumption on the kernels, for any $0 < \epsilon < 1$, let $n \geq M_1\eta/\epsilon$, then $\widehat{S}_0^u/n \leq \epsilon$ and $S_1/n \leq \epsilon$ with $S_1 = \mathbb{E}_{YY'}[\phi(Y,Y')]\mathbb{E}_{X_kX_k'}[\psi(X_k,X_k')]$. We have:

$$\begin{aligned}
&\mathbb{P}(|\widehat{S}_1^v - S_1| \geq 3\epsilon)\\
&\leq \mathbb{P}(|\frac{n-1}{n}\widehat{S}_1^u + \frac{1}{n}\widehat{S}_0^u - S_1| \geq 3\epsilon) \leq \mathbb{P}(|\frac{n-1}{n}\big(\widehat{S}_1^u - S_1\big)| \geq 2\epsilon - \frac{1}{n}S_1) \leq \mathbb{P}(|\widehat{S}_1^u - S_1| \geq \epsilon).
\end{aligned}$$

Now by Theorem 5.6.1.A of Serfling (1980),

$$\forall\epsilon > 0,\ \mathbb{P}(\widehat{S}_1^u - S_1 \geq \epsilon) \leq \exp(-c_1\lfloor n/2\rfloor\epsilon^2/\eta^2),$$

with $0 < c_1$ finite. The bound holds for deviations in the opposite direction: hence

$$\forall\epsilon > 0,\ \mathbb{P}(|\widehat{S}_1^v - S_1| \geq 3\epsilon) \leq 2\exp(-c_1\lfloor n/2\rfloor\epsilon^2/\eta^2).$$

As for $\widehat{S}_3^v$, for any $0 < \epsilon < 1$, let n such that $n \geq M_3\eta/\epsilon$, then $\frac{n-1}{n^2}\widehat{T}^u \leq \epsilon$ and $S_3/n \leq \epsilon$ with $S_3 = \mathbb{E}_{YX_k}[\mathbb{E}_{Y'}[\phi(Y,Y')]\mathbb{E}_{X_k'}[\psi(X_k,X_k')]]$. We obtain for a constant $K > 0$:

$$\begin{aligned}
&\mathbb{P}(|\widehat{S}_3^v - S_3| \geq 4\epsilon)\\
&\leq \mathbb{P}(|\frac{n-3}{n}\widehat{S}_3^u + \frac{n-1}{n^2}\widehat{T}^u + O(n^{-2}) - S_3| \geq 4\epsilon)\\
&\leq \mathbb{P}(|\frac{n-3}{n}\big(\widehat{S}_3^u - S_3\big)| \geq 3\epsilon - \frac{3}{n}S_3 - \frac{K}{n^2}) \leq \mathbb{P}(|\widehat{S}_3^u - S_3| \geq \epsilon).
\end{aligned}$$

Using the same argument for bounding $\widehat{S}_1^v$, we deduce for $0 < c_2$ finite

$$\mathbb{P}(|\widehat{S}_3^v - S_3| \geq 4\epsilon) \leq 2\exp(-c_2\lfloor n/3\rfloor\epsilon^2/\eta^2).$$

Finally let us treat $\widehat{S}_2^v$. For any $\epsilon > 0$, let n such that $n \geq M_4\eta/\epsilon$, then $\frac{(n-1)(n-2)}{n^3}\widehat{P}^u \leq \epsilon$ and $S_2/n \leq \epsilon$ where $S_2 = \mathbb{E}_{YY'X_kX'_k}[\phi(Y,Y')\psi(X_k,X'_k)]$. We obtain for any $\epsilon > 0$, for a constant $K' > 0$ and $c_3 > 0$:

$$\begin{aligned} & \mathbb{P}(|\widehat{S}_2^v - S_2| \geq 4\epsilon) \\ \leq \;& \mathbb{P}(|\frac{n-6}{n}\widehat{S}_2^u + \frac{(n-1)(n-2)}{n^3}\widehat{P}^u + O(n^{-2}) - S_3| \geq 4\epsilon) \\ \leq \;& \mathbb{P}(|\frac{n-6}{n}\big(\widehat{S}_3^u - S_3\big)| \geq 3\epsilon - \frac{6}{n}S_3 - \frac{K'}{n^2}) \leq \mathbb{P}(|\widehat{S}_3^u - S_3| \geq \epsilon) \leq 2\exp(-c_3\lfloor n/4\rfloor\epsilon^2/\eta^2). \end{aligned}$$

We then obtain for $c > 0$ a finite constant:

$$\begin{aligned} & \forall \epsilon > 0,\; \mathbb{P}(|\widehat{\text{HSIC}}_v(\mathbf{Y},\mathbf{X}_k) - \text{HSIC}(Y,X_k)| \geq \epsilon) \\ & \quad\leq\; \mathbb{P}(|\widehat{S}_1^v + \widehat{S}_2^v - 2\widehat{S}_3^v - (S_1 + S_2 - 2S_3)| \geq \epsilon) \\ & \quad\leq\; \mathbb{P}(|\widehat{S}_1^v - S_1| \geq \epsilon/3) + \mathbb{P}(|\widehat{S}_2^v - S_2| \geq \epsilon/3) + \mathbb{P}(|\widehat{S}_3^v - S_3| \geq \epsilon/6) \leq O(\exp(-cn\epsilon^2)), \end{aligned}$$

with $C > 0$ a finite constant. Thus the bound over $\max\limits_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k|$ follows. The sure screening property can then be deduced as in Theorem 3.1.

□

Proof of Lemma 4.1

Proof. We proceed in the same spirit as in Barber and Candès (2015). Let the function $\psi : \mathbb{R}^{2p+1} \longrightarrow \mathbb{R}^p$ defined as $\psi(\mathbf{X},\widetilde{\mathbf{X}},\mathbf{Y}) = (\widehat{W}_1,\cdots,\widehat{W}_p)^\top$. For any $j \in \mathcal{A} \subset \{1,\cdots,p\}$, we denote by $(\mathbf{X},\widetilde{\mathbf{X}})_{(j)}$ the vector deduced from swapping the entries $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$ in $(\mathbf{X},\widetilde{\mathbf{X}})$. The quantity $(\mathbf{X},\widetilde{\mathbf{X}})_{\text{swap}(\mathcal{A})}$ is the vector version for all $j \in \mathcal{A}$. Due to the anti-symmetry property of $\widehat{W}_j$,

$$\widehat{W}_{\text{swap}(\mathcal{A})} = (\epsilon_1\widehat{W}_1,\cdots,\epsilon_p\widehat{W}_p), \;\text{ with }\; \epsilon_j = \mathbf{1}(j \notin \mathcal{A}) - \mathbf{1}(j \in \mathcal{A}).$$

Using the argument (i) of Lemma 1 of Liu et al. (2020), $(\mathbf{Y},(\mathbf{X},\widetilde{\mathbf{X}})) \overset{d}{=} (\mathbf{Y},(\mathbf{X},\widetilde{\mathbf{X}})_{\text{swap}(\mathcal{A})})$ for any $\mathcal{A} = \{1,\cdots,p\}\setminus\mathcal{S}$: point (i) of Lemma 1 of these authors established that for any $j \in \mathcal{S}^c$, then $(\mathbf{Y},\mathbf{X}_j) = (\mathbf{Y},\widetilde{\mathbf{X}}_j)$ in distribution, that is $W_j = 0$. Now defining $\mathcal{S} = \{j : \epsilon_j = -1\}$ the set of inactive features, we deduce

$$(\widehat{W}_1,\cdots,\widehat{W}_p)^\top_{\text{swap}(\mathcal{A})} = \psi((\mathbf{X},\widetilde{\mathbf{X}})_{\text{swap}(\mathcal{A})},\mathbf{Y}) \overset{d}{=} \psi(\mathbf{X},\widetilde{\mathbf{X}},\mathbf{Y}).$$

□

Proof of Theorem 4.2

Proof. Let us define $\mathcal{G} = \mathcal{S} \cap \widehat{\mathcal{S}}_0$ and $\mathcal{G}^c = \mathcal{S}^c \cap \widehat{\mathcal{S}}_0$. We proceed in the same spirit as in Lemma 1 of the supplement of Barber and Candès (2015) or Theorem 4 of Liu et al. (2020). We take $\widehat{\mathcal{S}}_0 = \{1,\cdots,s_0\}$ and $|\widehat{W}_1| \geq \cdots \geq |\widehat{W}_{s_0}| > 0$ and omit the conditioning with respect to $\mathcal{E}$. Then, we have

$$\begin{aligned} & \mathbb{E}[\frac{\text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; j \in \widehat{\mathcal{S}}(\alpha)}{\text{card}(j:\; j \in \widehat{\mathcal{S}}(\alpha)) \vee 1}] = \mathbb{E}[\frac{\text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; \widehat{W}_j \geq T(\alpha)}{\text{card}(j:\; \widehat{W}_j \geq T(\alpha)) \vee 1}] \\ =\;& \mathbb{E}[\frac{\text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; \widehat{W}_j \geq T(\alpha)}{1 + \text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; \widehat{W}_j \leq -T(\alpha))} \times \frac{1 + \text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; \widehat{W}_j \leq -T(\alpha))}{\text{card}(j:\; \widehat{W}_j \geq T(\alpha)) \vee 1}] \\ \leq\;& \mathbb{E}[\frac{\text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; j \in \widehat{W}_j \geq T(\alpha)}{1 + \text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; \widehat{W}_j \leq -T(\alpha))} \times \frac{1 + \text{card}(j: \widehat{W}_j \leq -T(\alpha))}{\text{card}(j:\; \widehat{W}_j \geq T(\alpha)) \vee 1}] \\ \leq\;& \alpha\;\; \mathbb{E}[\frac{\text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; j \in \widehat{W}_j \geq T(\alpha)}{1 + \text{card}(j:\; j \in \mathcal{G}^c \;\text{ and }\; \widehat{W}_j \leq -T(\alpha))}], \end{aligned}$$

since $\big\{j :\; j \in \mathcal{G}^c \cap \{\widehat{W}_j \leq -T(\alpha)\}\big\} \subseteq \big\{j : \widehat{W}_j \leq -T(\alpha)\big\}$ and by definition of the threshold $T(\alpha)$, $\frac{1+\text{card}(j:\widehat{W}_j \leq -T(\alpha))}{\text{card}(j:\, \widehat{W}_j \geq T(\alpha))\vee 1} \leq \alpha$. The threshold $T(\alpha)$ is such that $|\widehat{W}_1| \geq |\widehat{W}_2| \geq \cdots \geq |\widehat{W}_{s_0}| \geq T(\alpha) > |\widehat{W}_{s_0+1}| \geq \cdots$ and can be seen as a stopping time: $T(\alpha)$ can be set by choosing the smallest t value such that $|\widehat{W}_{s_0}|=0$; then choose a larger t such that $|\widehat{W}_{s_0-1}|=0$; and so on, such that $T(\alpha)$ corresponds to the t value satisfying $\frac{1+\mathrm{card}(j:\widehat{W}_j\le -t)}{\mathrm{card}(j:\,\widehat{W}_j\ge t)\vee 1}\le\alpha$. In the same spirit as Lemma 1 of the supplement of Barber and Candès (2015), we show that $T(\alpha)$ is a stopping time for the process $V^+(t)/(1+V^-(t))$ with $V^{\pm}(t)=\mathrm{card}(j:\ j\in\mathcal{G}^c,\ |\widehat{W}_j|\ge t,\ \mathrm{sgn}(\widehat{W}_j)=\pm 1)$. We consider the process $M(k)=\frac{V^+(k)}{1+V^-(k)}, k=s_0,s_0-1,\cdots,1,0$, with

$$V^+(k)=\mathrm{card}(j:j\in\mathcal{G}^c,1\le j\le k,\widehat{W}_j>0),\ V^-(k)=\mathrm{card}(j:j\in\mathcal{G}^c,1\le j\le k,\widehat{W}_j\le 0),$$

and prove $M(k)$ is a supermartingale for the backward filtration $\mathcal{F}_k=\sigma(V^{\pm}(s),B_s:=\mathbf{1}(s\in\mathcal{G}),s\ge k)$ so that $\mathcal{F}_{s_0}\subset\mathcal{F}_{s_0-1}\subset\cdots\subset\mathcal{F}_1$. The latter informs whether $k\in\mathcal{G}$ or $k\in\mathcal{G}^c$. If $k\in\mathcal{G}$, then $V^{\pm}(k)=V^{\pm}(k-1)$ so that $M(k)=M(k-1)$. Now if $k\in\mathcal{G}^c$, then

$$M(k-1)=\frac{V^+(k)-I_k}{1+V^-(k)-(1-I_k)}=\frac{V^+(k)-I_k}{(V^-(k)+I_k)\vee 1},\ \ I_k=\mathbf{1}(\widehat{W}_k>0).$$

Since the true inactive features are uniformly distributed and by the exchangeability property of the inactive features from Lemma 4.1, then $\mathbb{P}(I_k=1|\mathcal{F}_k)=\frac{V^+(k)}{V^+(k)+V^-(k)}$. As a consequence, for any $k\in\mathcal{G}^c$, we deduce

$$\begin{aligned}\mathbb{E}[M(k-1)|\mathcal{F}_k]&=\mathbb{E}[\frac{V^+(k)-I_k}{(V^-(k)+I_k)\vee 1}|\mathcal{F}_k]\\
&=\ \mathbb{P}(I_k=1|\mathcal{F}_k)\big(\frac{V^+(k)-1}{(V^-(k)+1)\vee 1}\big)+(1-\mathbb{P}(I_k=1|\mathcal{F}_k))\big(\frac{V^+(k)}{V^-(k)\vee 1}\big)\\
&=\ \frac{V^+(k)}{V^+(k)+V^-(k)}\big(\frac{V^+(k)-1}{V^-(k)+1}\big)+\frac{V^-(k)}{V^+(k)+V^-(k)}\big(\frac{V^+(k)}{V^-(k)\vee 1}\big)\\
&=\ \big(V^+(k)-1\big)\mathbf{1}(V^-(k)=0)+\big(\frac{V^+(k)}{1+V^-(k)}\big)\mathbf{1}(V^-(k)>0).\end{aligned}$$

As a consequence

$$\mathbb{E}[M(k-1)|\mathcal{F}_k]=\begin{cases}M(k), & k\in\mathcal{G},\\ M(k), & k\in\mathcal{G}^c,\ \text{ and }\ V^-(k)>0,\\ M(k)-1, & k\in\mathcal{G}^c,\ \text{ and }\ V^-(k)=0,\end{cases}$$

which implies that $(M(k))_k$ is a supermartingale with respect to $\mathcal{F}_k$. Now $T(\alpha)$ is a stopping time for the filtration $\mathcal{F}_k$ as $\{T(\alpha)\ge k\}\in\mathcal{F}_k$, so that $\mathbb{E}[M(k_{T(\alpha)})|\mathcal{F}_{k_{s_0}}]\le M(k_{s_0})$ by the optional stopping time Theorem. Taking the expectation on both sides, we deduce

$$\mathbb{E}[M(k_{T(\alpha)})]\le\mathbb{E}[M(k_{s_0})]=\mathbb{E}[\frac{\mathrm{card}(j:\ j\in\mathcal{G}^c,\ \widehat{W}_j>0)}{1+\mathrm{card}(j:\ j\in\mathcal{G}^c,\ \widehat{W}_j\le 0)}].$$

Let $N=\mathrm{card}(j:\ j\in\mathcal{G}^c)$, setting $X=\mathrm{card}(j:\ j\in\mathcal{G}^c,\ \widehat{W}_j>0)$, it follows $X\sim\mathcal{B}(N,0.5)$ a Binomial distribution by independence of the inactive indices. We hence deduce

$$\begin{aligned}\mathbb{E}[M(k_{s_0})]&=\mathbb{E}[\frac{X}{1+N-X}]\\
&=\ \sum_{r=1}^{N}\mathbb{P}(X=r)\times\frac{r}{1+N-r}=\sum_{r=1}^{N}\binom{N}{r}\big(\frac{1}{2}\big)^r\big(\frac{1}{2}\big)^{N-r}\frac{r}{1+N-r}=\sum_{r=0}^{N-1}\binom{N}{r}\big(\frac{1}{2}\big)^{r+1}\big(\frac{1}{2}\big)^{N-r-1}\le 1.\end{aligned}$$

As a consequence,

$$\mathbb{E}[\frac{\mathrm{card}(j:\ j\in\mathcal{G}^c\ \text{ and }\ j\in\widehat{W}_j\ge T(\alpha))}{1+\mathrm{card}(j:\ j\in\mathcal{G}^c\ \text{ and }\ \widehat{W}_j\le -T(\alpha))}]\le\mathbb{E}[M(k_{T(\alpha)})]\le 1.$$

Now $\widehat{\mathcal{S}}(\alpha)\subseteq\widehat{\mathcal{S}}_0$ implies $\mathrm{card}(j:j\in\mathcal{S}^c,j\in\widehat{W}_j\ge T(\alpha))=\mathrm{card}(j:j\in\mathcal{G}^c,j\in\widehat{W}_j\ge T(\alpha))$, we obtain the desired control. □

C Implementation procedures

C.1 Python package

We made all of our code open source on Github. The code consists of Python 3 files implementing Scikit-Learn (Buitinck et al., 2013) like object; in particular it has a fit attribute which consists of applying the knockoff algorithm. All the association measures with different kernels are implemented when applicable. The knockoff variable, carried out according to the equicorrelated construction, and the Projection Correlation's implementation were written in Python 2 and were made available by Liu et al. (2020). Distance Correlation's implementation was taken from the *dcor* pypi package.

In Figure 1, we report the computation and memory consumption of all association measures. During knockoff we compute $\mathrm{D}(Y, X_k)$, therefore only the size of these vectors, i.e n influences the processing time, the number of features p multiplies the processing time by p. In particular, we notice that PC is by far the worst in both aspect.

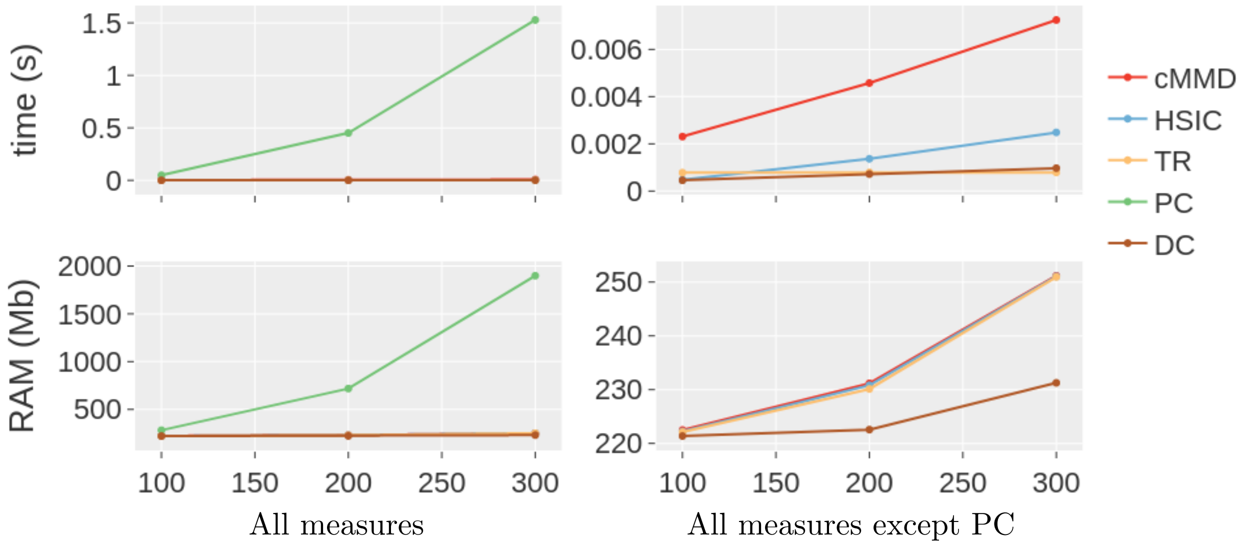

Figure 1: Monitoring CPU time and memory consumption for the five main association measures. We have the sample size on the x-axis.

C.2 Pipeline

In addition, we provide Nextflow (Di Tommaso et al., 2017) scripts to reproduce our results and plots from a single file. Nextflow can be paired with Singularity (Kurtzer et al., 2017) in order to containerise the process to make it easily adaptable and configurable to most platform. The False Detection Rate and Minimum model size plots were generated from this single script. The code is split into three processes: data generation, knockoff fit and plotting. Similarly, the code to reproduce the applications on MNIST and TCGA are also available in a very similar format. Hence this allows any external user to easily adapt the code to new data, by modifying the first processes, and to new association measures by modifying the second process.

D Figures

In Figures 2 to 10 we show the minimum size model for each DGP introduced in the main paper except for some cases where the high complexity of PC can't be experimented with. Specifically we were not able to perform the 200 replicas for PC when $n = 1000$ and $p = 5000$. We display in shades of blue the usage of different kernels for HSIC and in shades of red those for cMMD. In a dark line, we display the median and in dash the mean of the distribution. The 5% and 95% are given by the whiskers. To help differentiate and compare distributions we also display in a box the 3rd quartile value.

In Figures 11 to 19 we plot FDR vs α to show that the knockoff procedure bounds the FDR.

In Figures 20 to 28 we plot FDR vs α for the association measures based on a kernel, to show that the kernel knockoff procedure bounds the FDR.

In Figure 29, we plot the empirical probability of returning an empty set for each measure.

In Figure 30 and Figure 31 respectively, we plot the minimum model size, FDR control and the percentage of returning the empty set for increasing values of n_0 with respect to the row, which is set to 10% (the first row), 30%, 50%, 70% and 90% (the last row) of n for DGP 2a and 2b respectively.

In Figures 32 and 33 we show the analyses for the applications to two real-world datasets: MNIST and breast cancer gene expression.

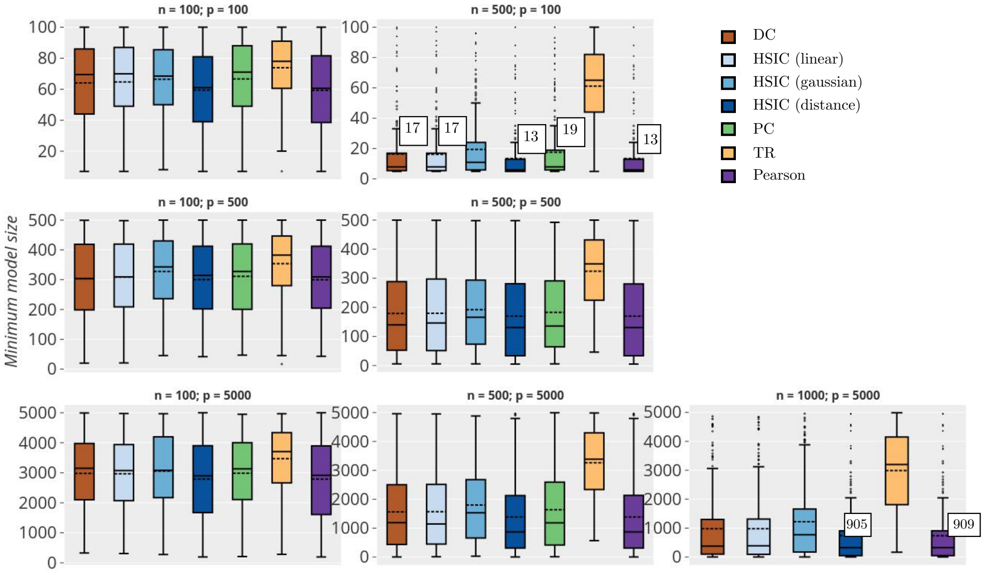

Figure 2: Minimum model size for DGP 1.a. For the sake of comparison, we display the third quartile of the distribution.

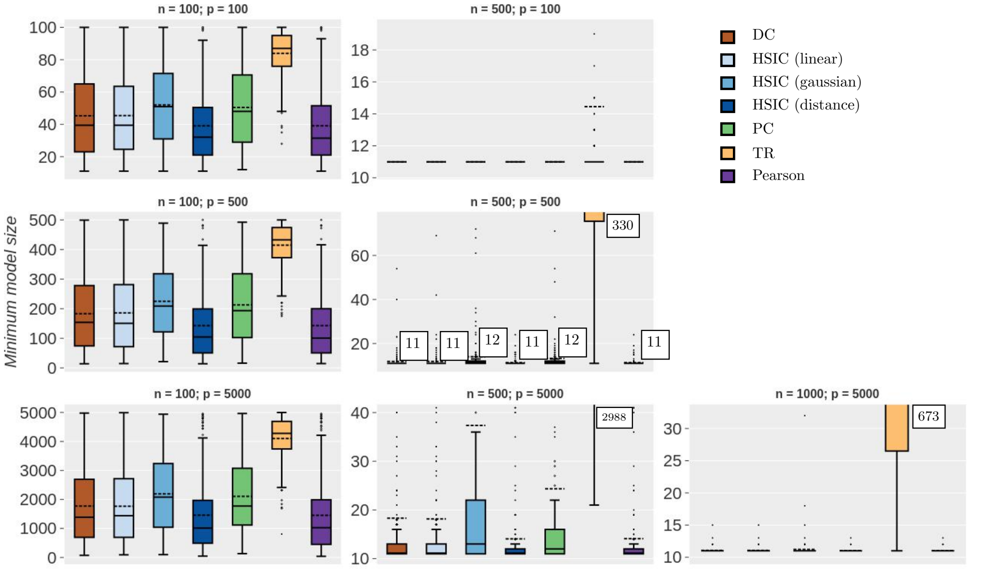

Figure 3: Minimum model size for DGP 1.b. For the sake of comparison, we display the third quartile of the distribution.

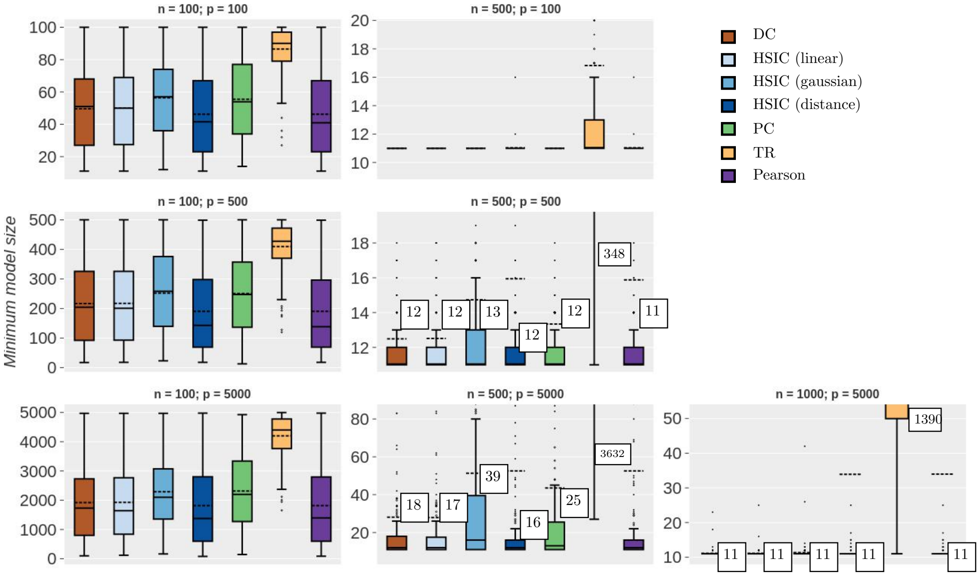

Figure 4: Minimum model size for DGP 1.c. For the sake of comparison, we display the third quartile of the distribution.

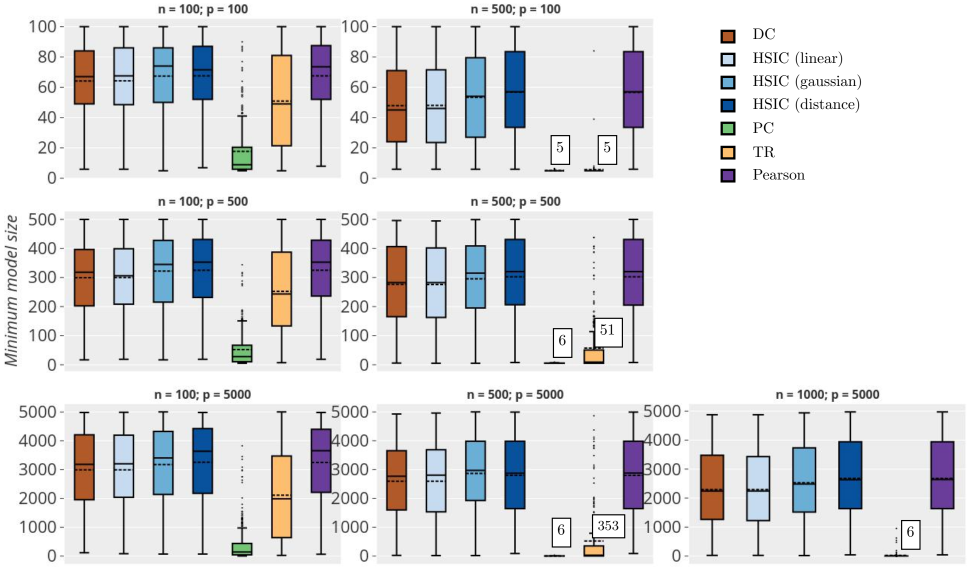

Figure 5: Minimum model size for DGP 2.a. For the sake of comparison, we display the third quartile of the distribution.

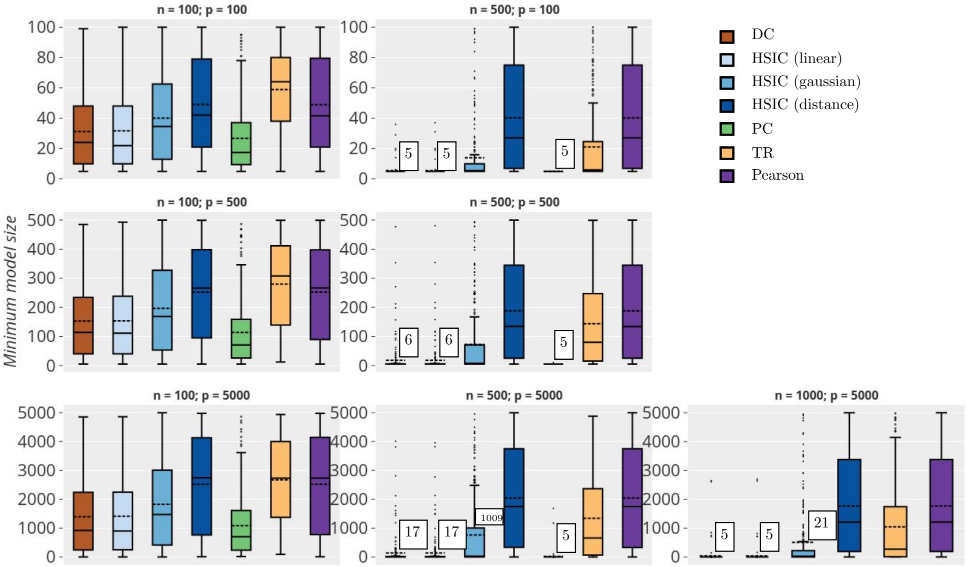

Figure 6: Minimum model size for DGP 2.b. For the sake of comparison, we display the third quartile of the distribution.

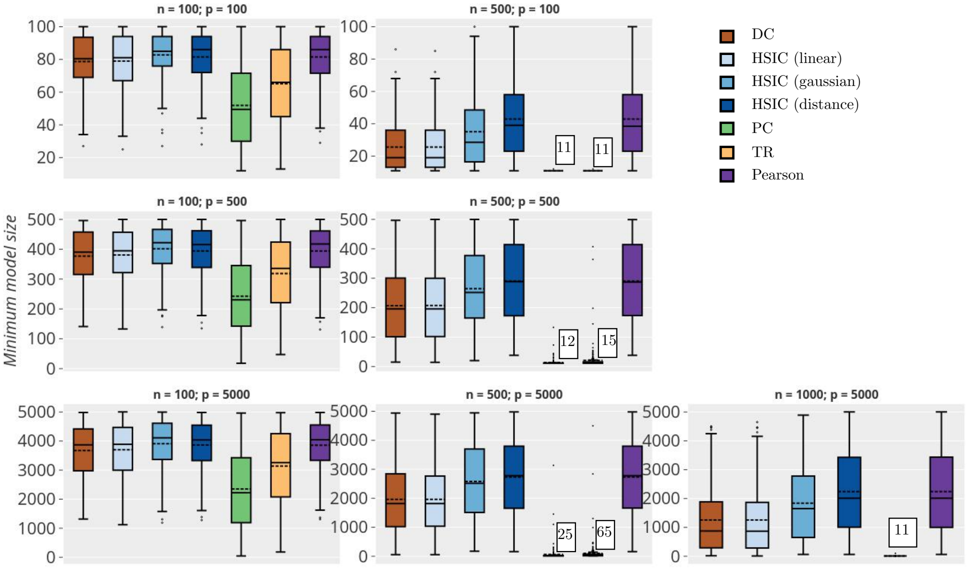

Figure 7: Minimum model size for DGP 2.c. For the sake of comparison, we display the third quartile of the distribution.

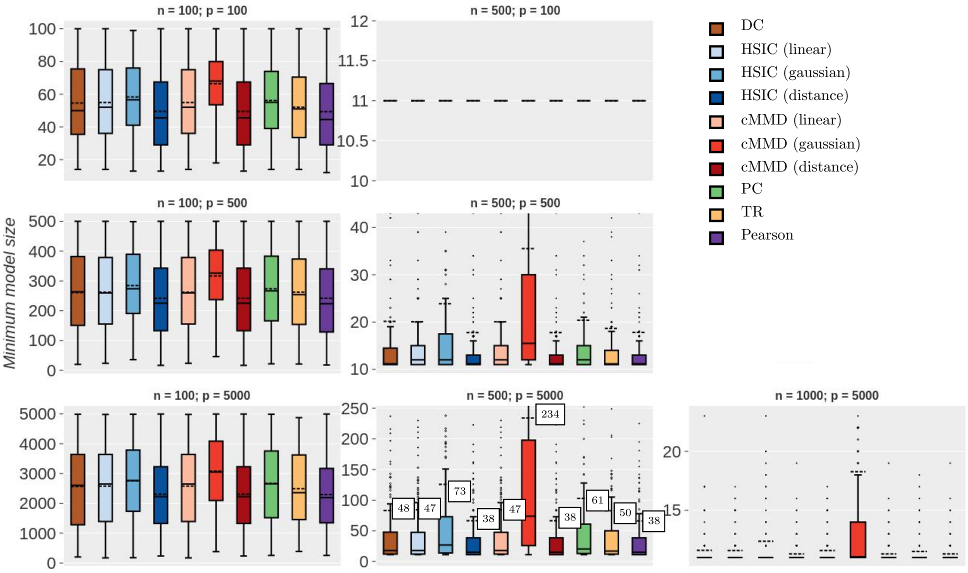

Figure 8: Minimum model size for DGP 3.a. For the sake of comparison, we display the third quartile of the distribution.

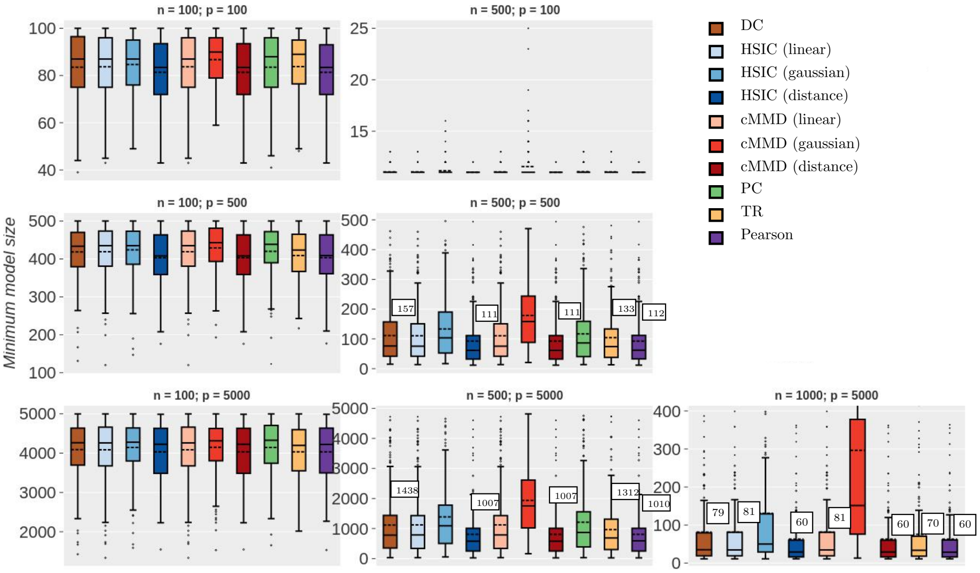

Figure 9: Minimum model size for DGP 3.b. For the sake of comparison, we display the third quartile of the distribution.

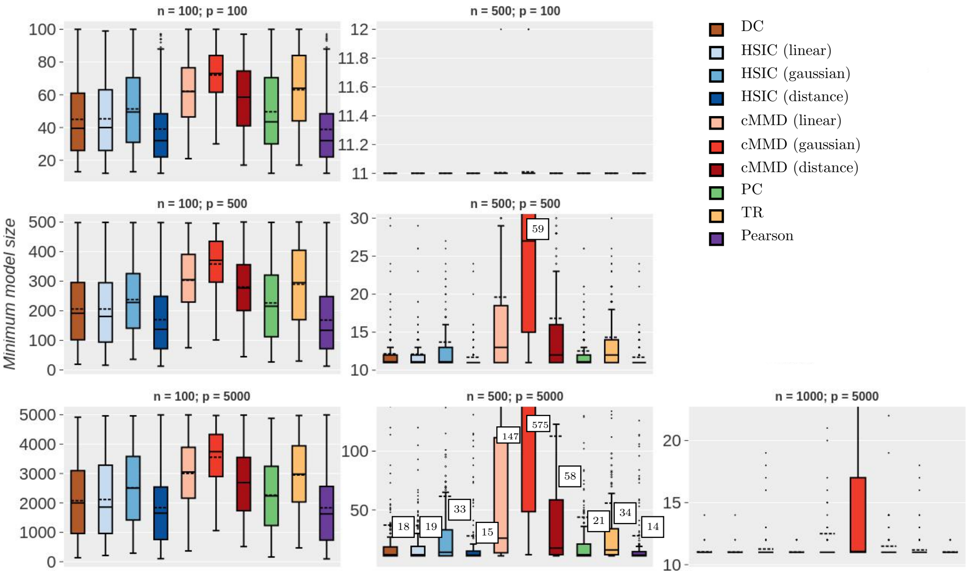

Figure 10: Minimum model size for DGP 3.c. For the sake of comparison, we display the third quartile of the distribution.

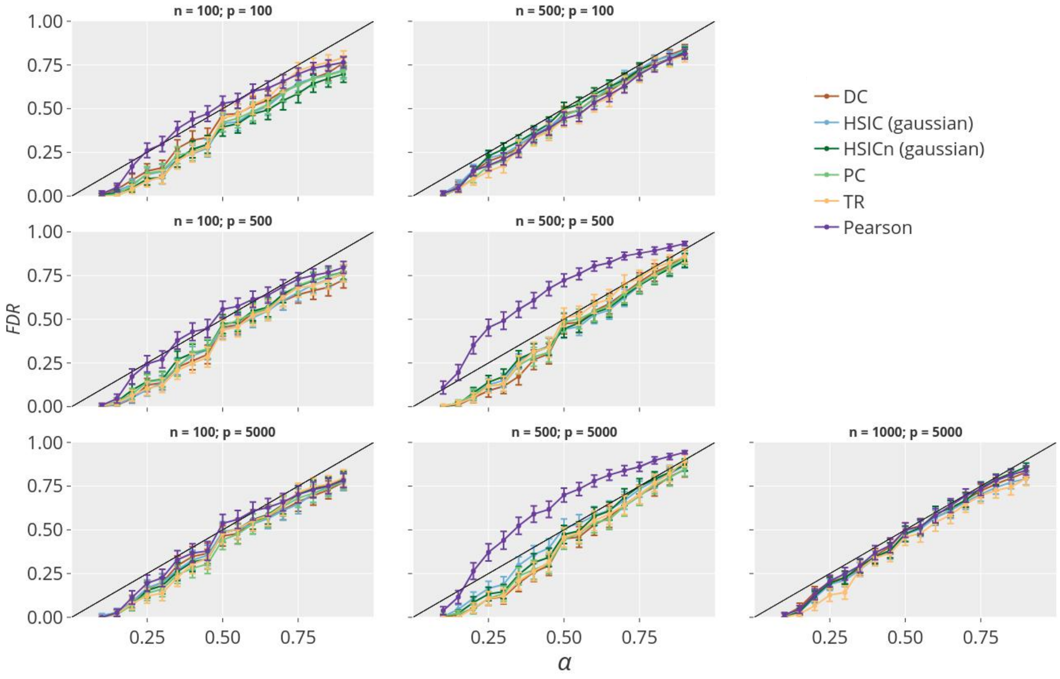

Figure 11: FDR control for DGP 1.a

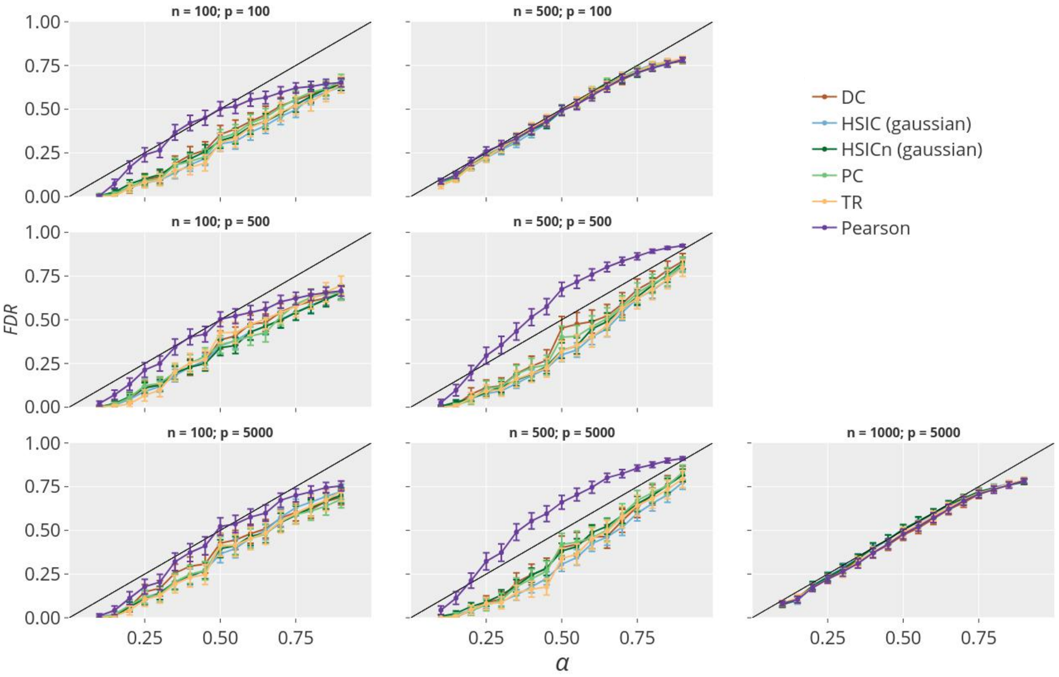

Figure 12: FDR control for DGP 1.b

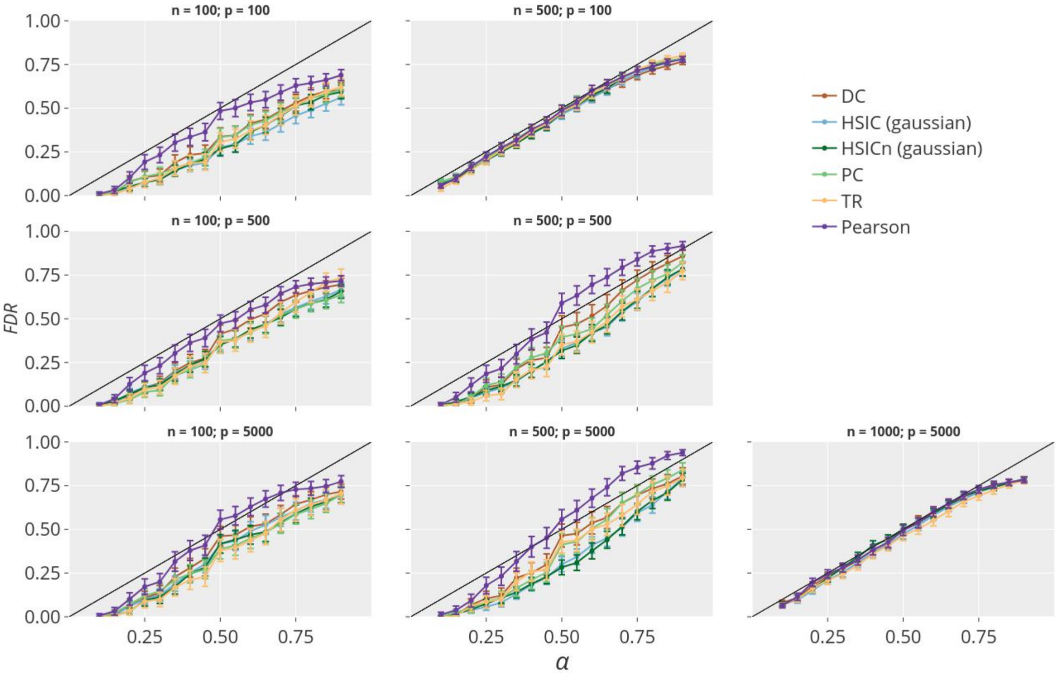

Figure 13: FDR control for DGP 1.c

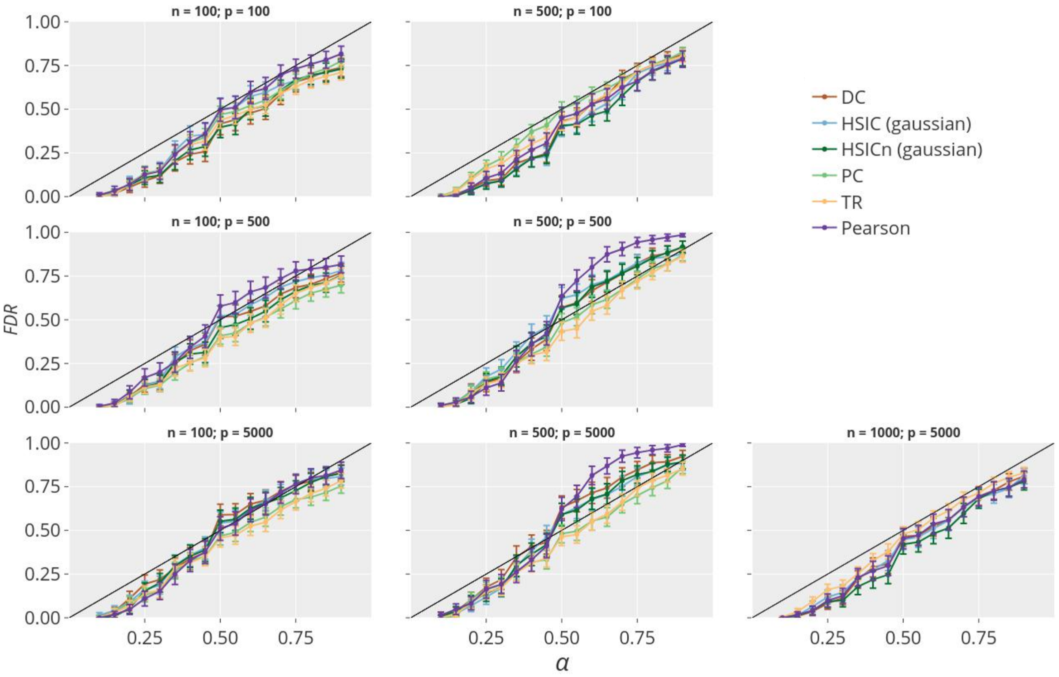

Figure 14: FDR control for DGP 2.a

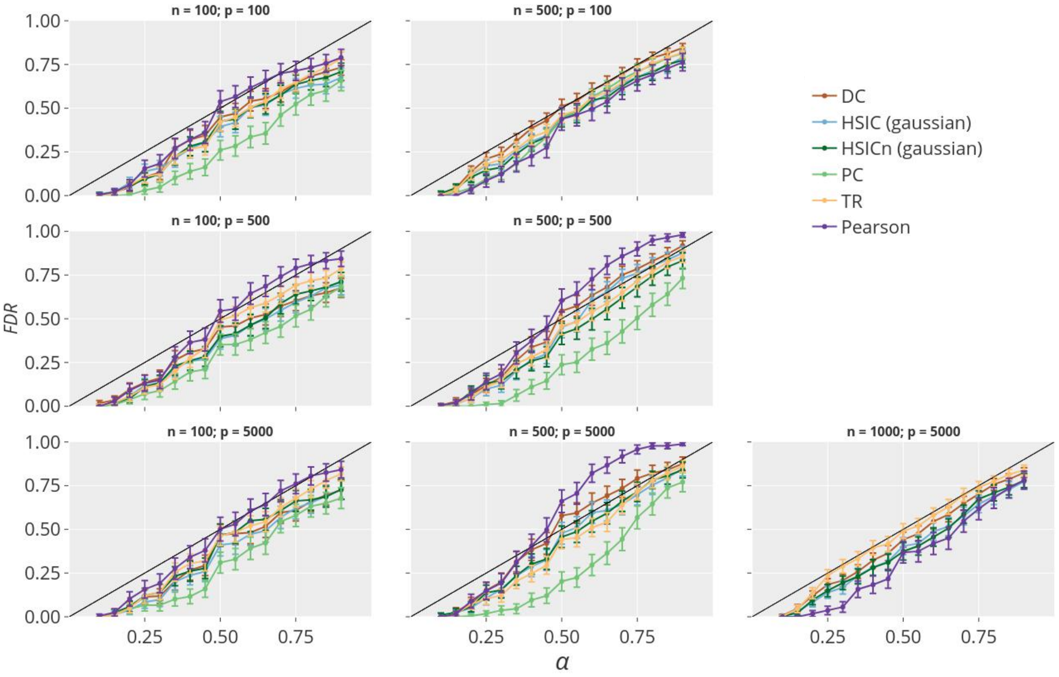

Figure 15: FDR control for DGP 2.b

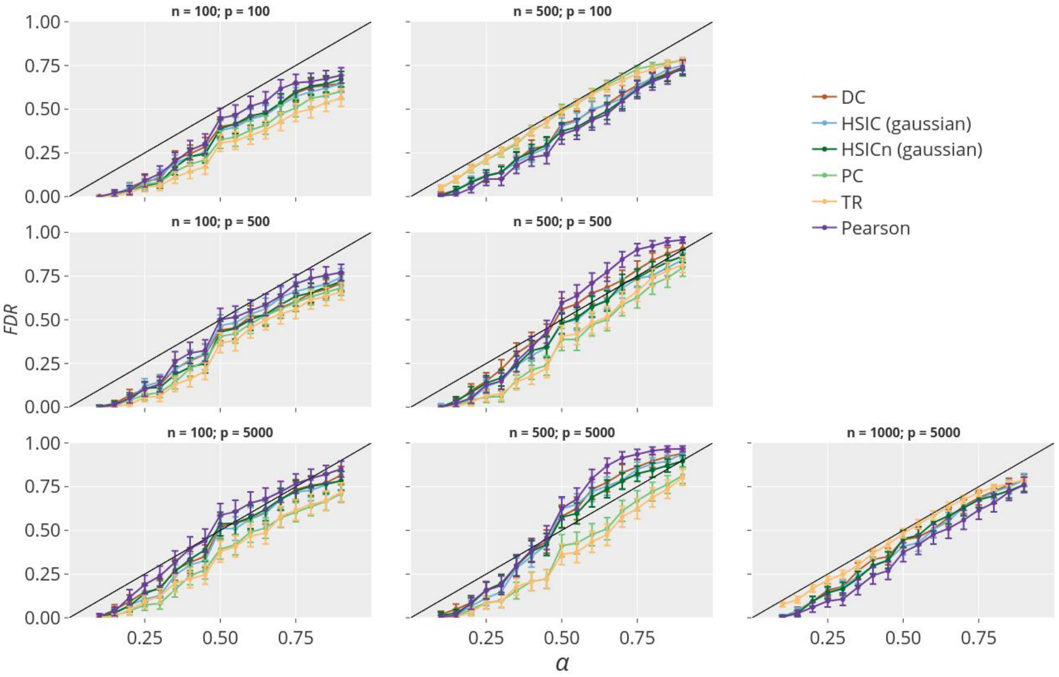

Figure 16: FDR control for DGP 2.c

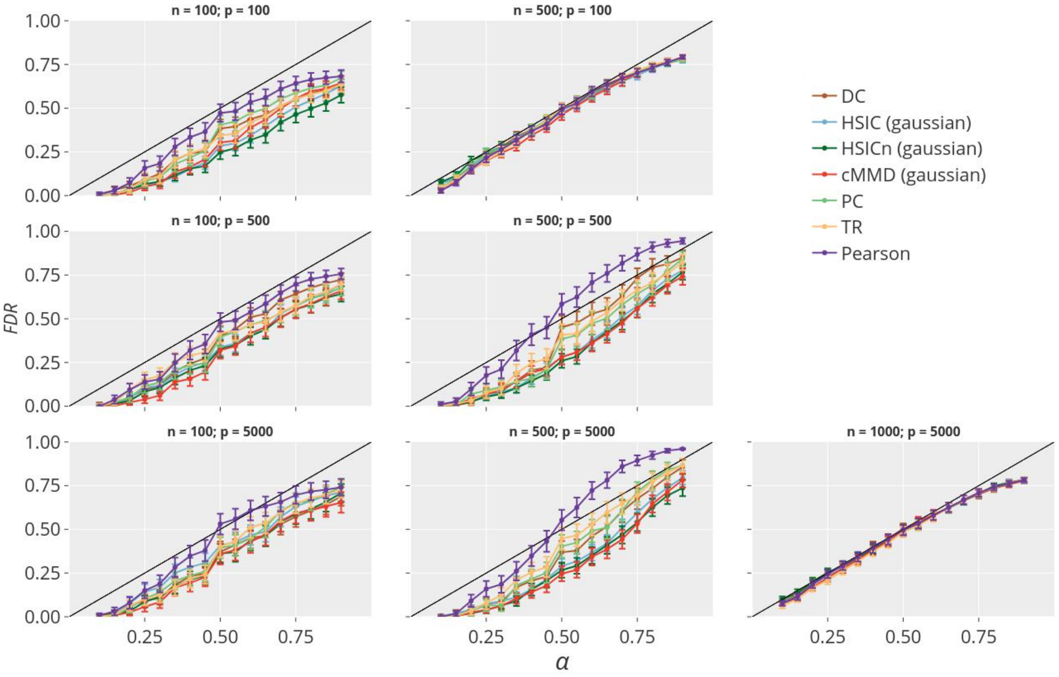

Figure 17: FDR control for DGP 3.a

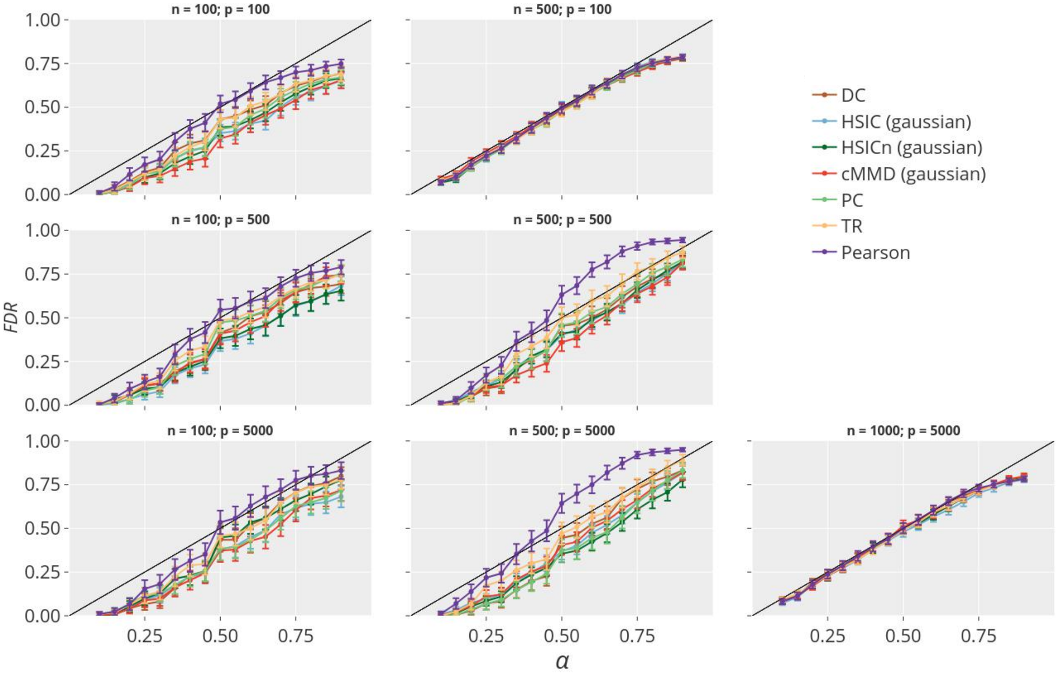

Figure 18: FDR control for DGP 3.b

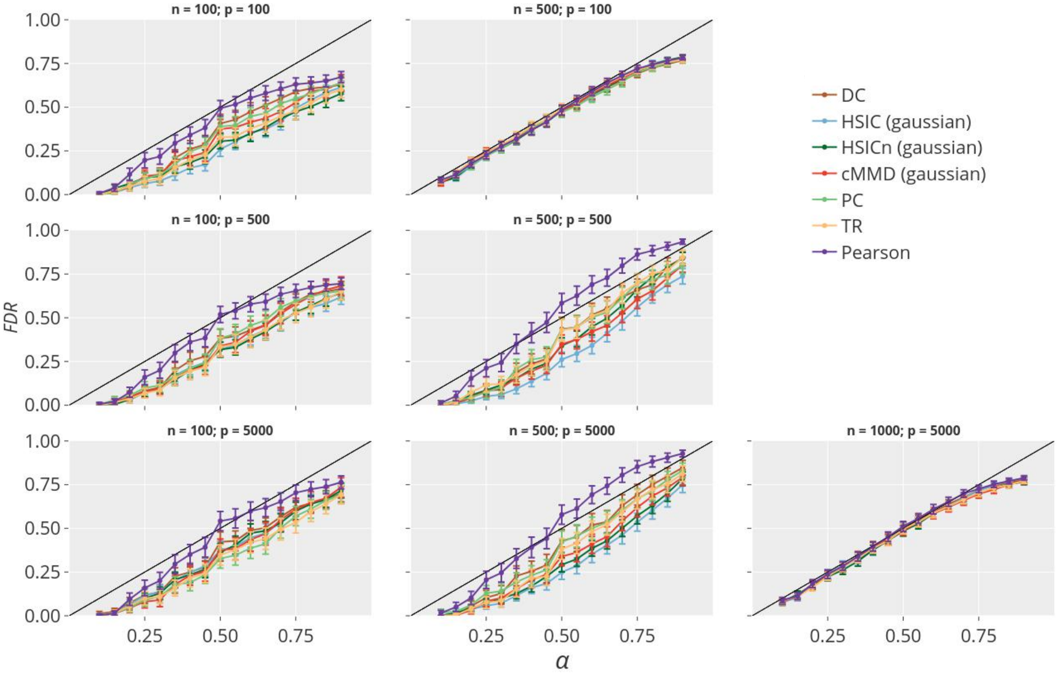

Figure 19: FDR control for DGP 3.c

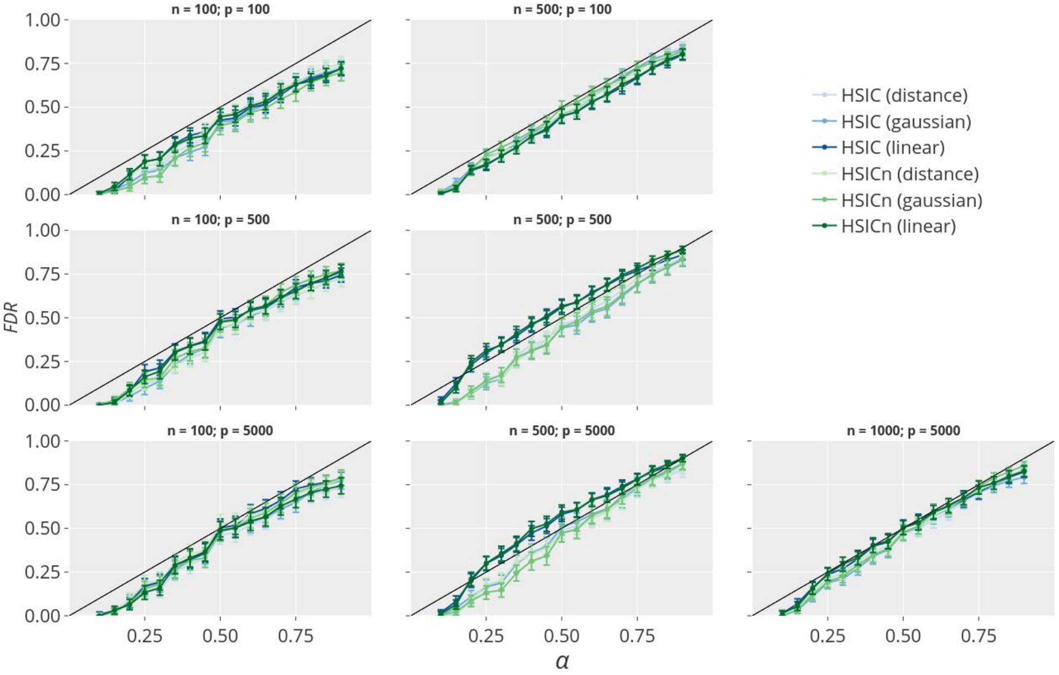

Figure 20: FDR control only for the kernel association measures for DGP 1.a

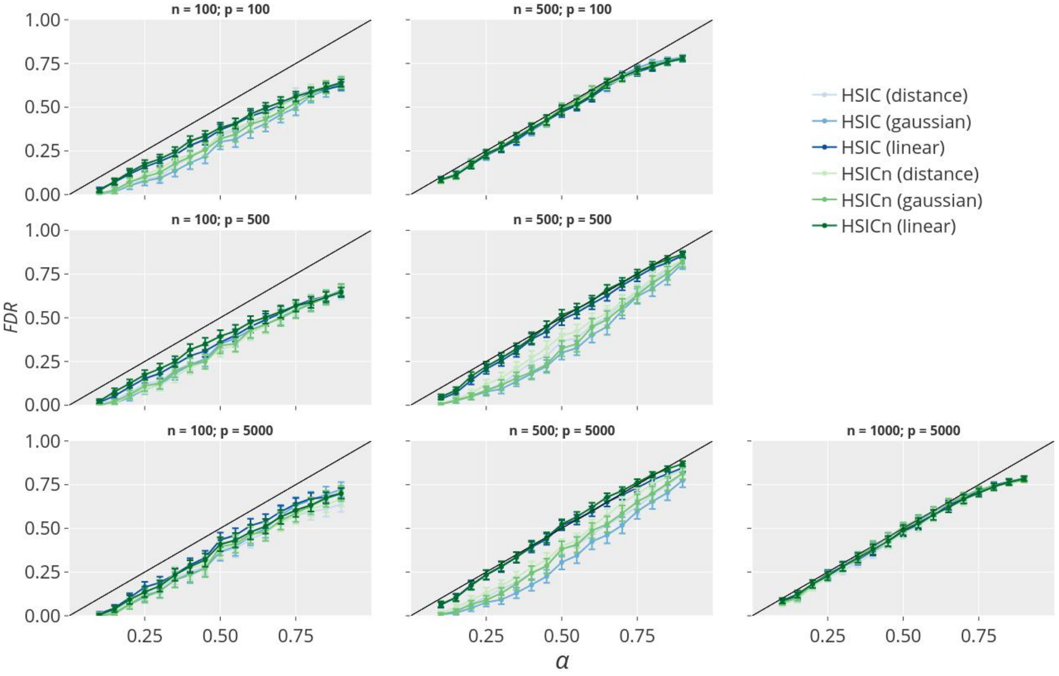

Figure 21: FDR control only for the kernel association measures for DGP 1.b

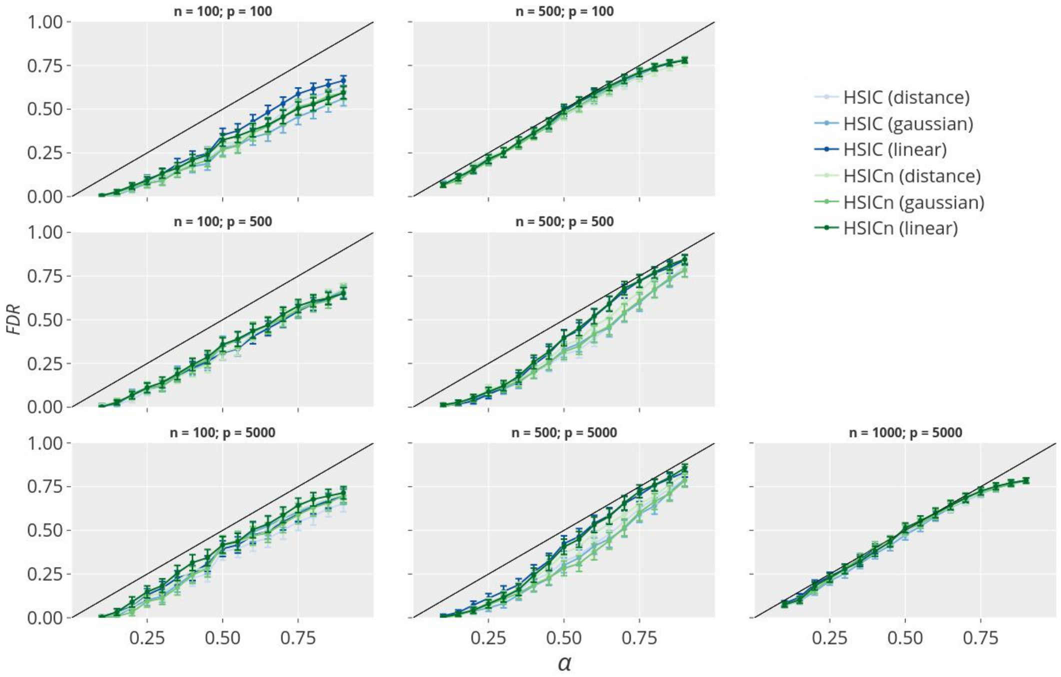

Figure 22: FDR control only for the kernel association measures for DGP 1.c

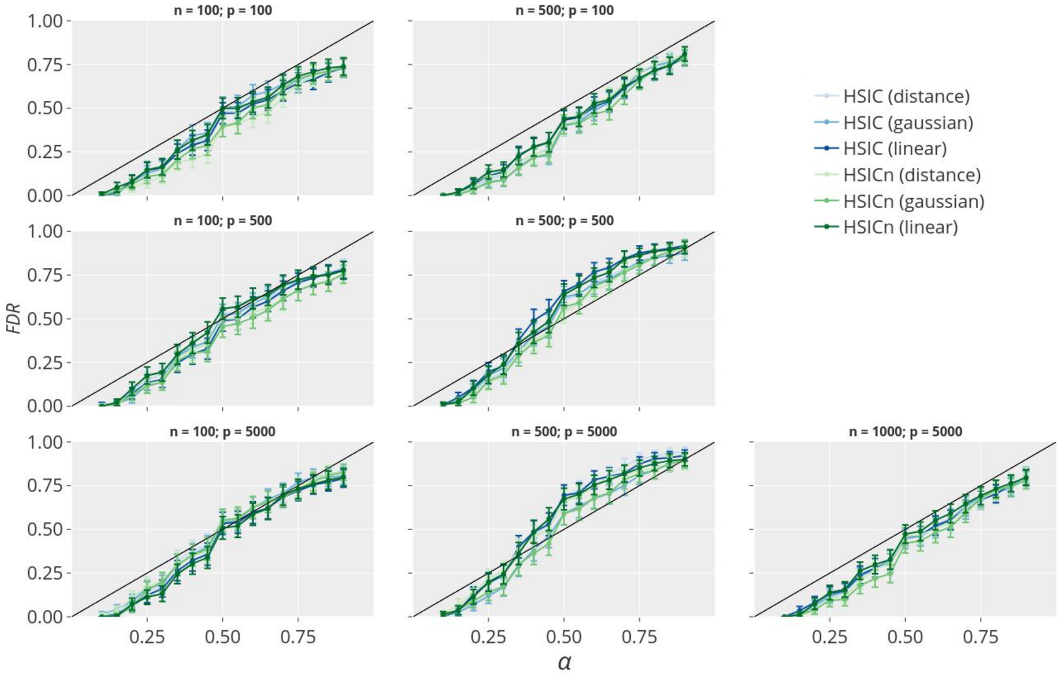

Figure 23: FDR control only for the kernel association measures for DGP 2.a

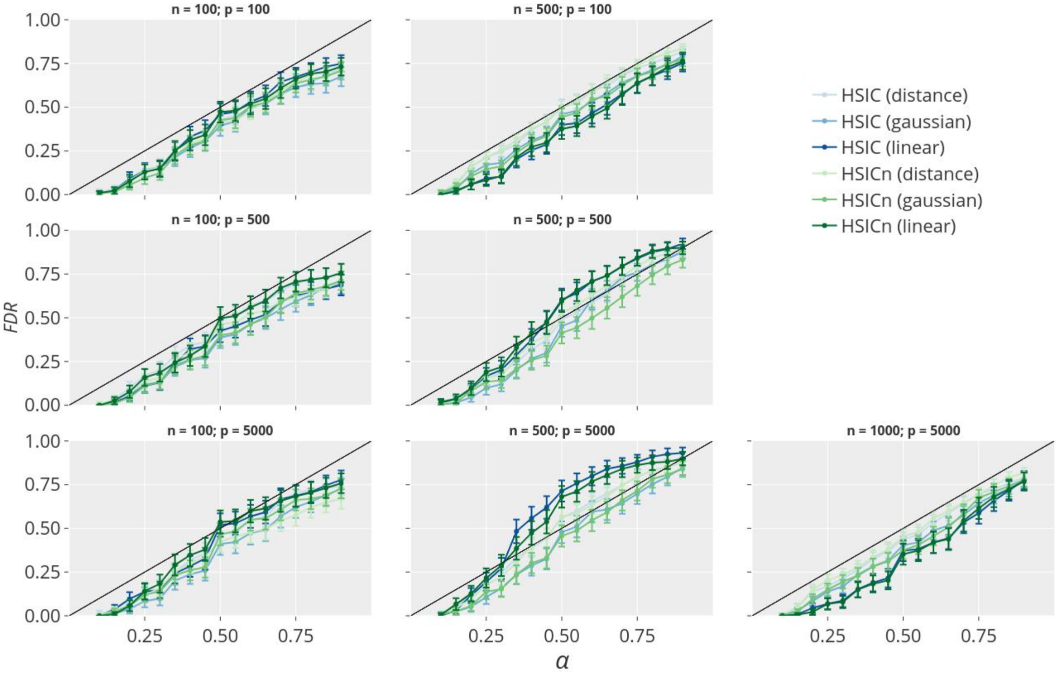

Figure 24: FDR control only for the kernel association measures for DGP 2.b

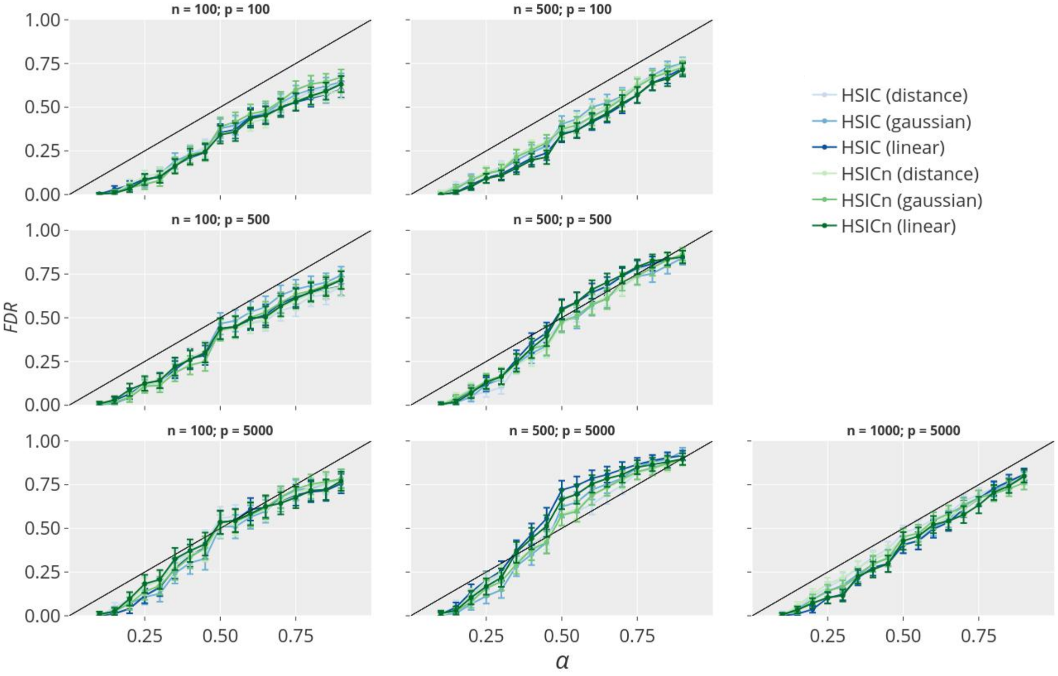

Figure 25: FDR control only for the kernel association measures for DGP 2.c

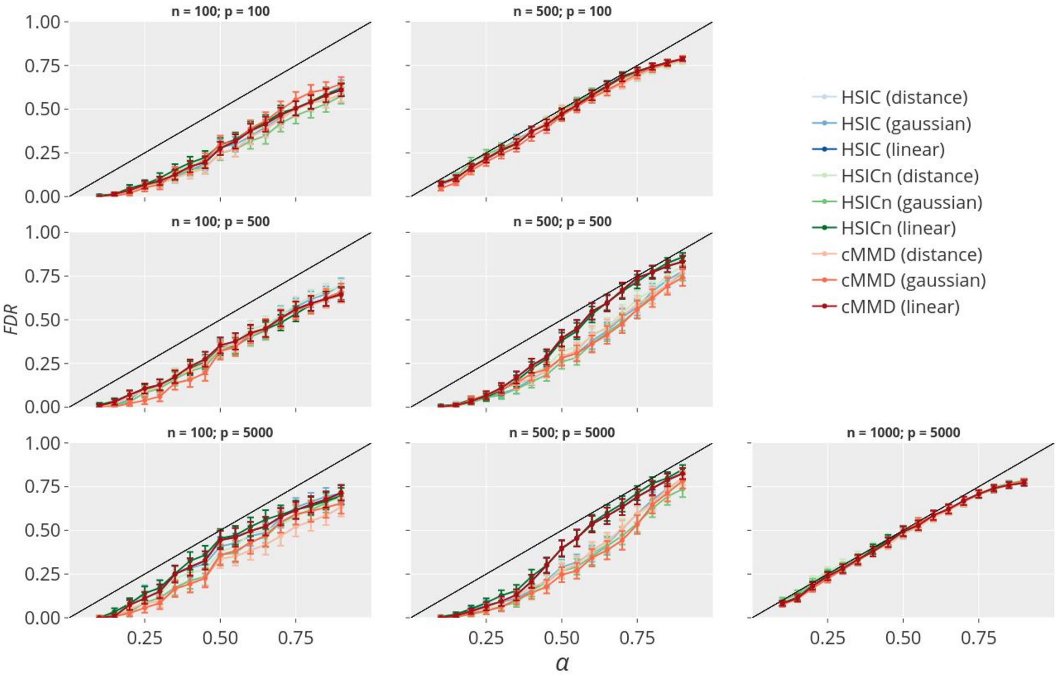

Figure 26: FDR control only for the kernel association measures for DGP 3.a

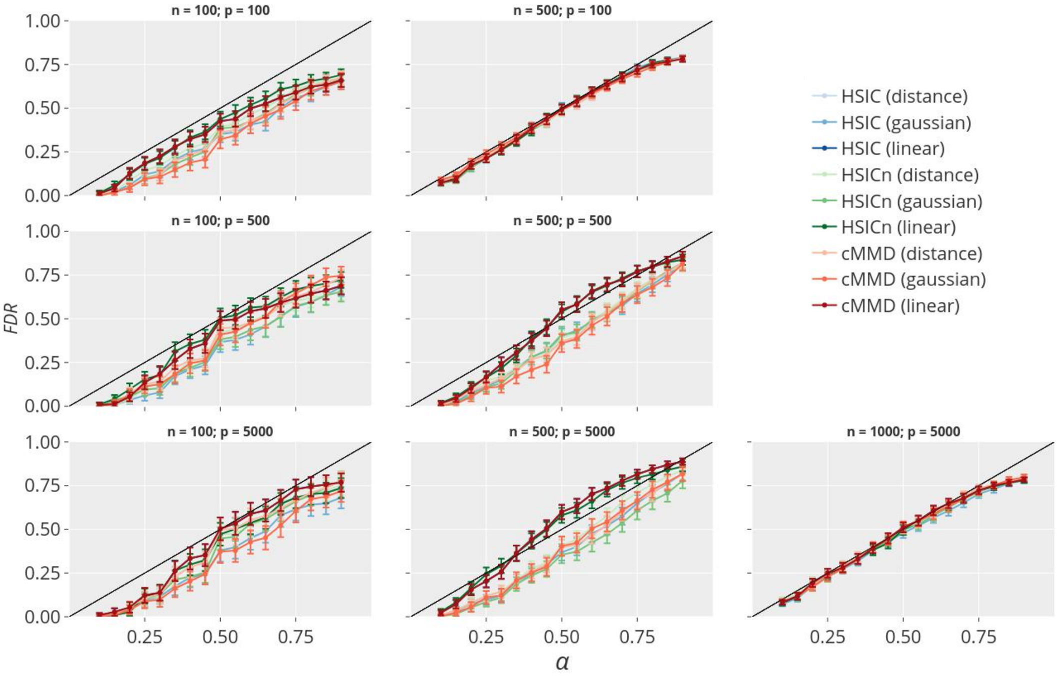

Figure 27: FDR control only for the kernel association measures for DGP 3.b

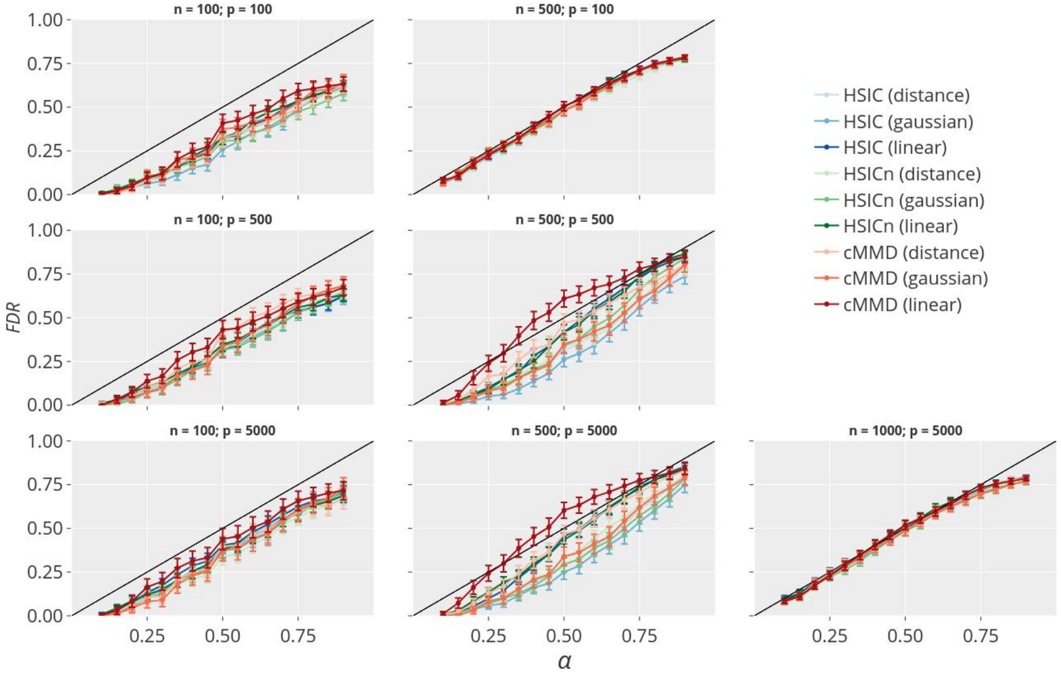

Figure 28: FDR control only for the kernel association measures for DGP 3.c

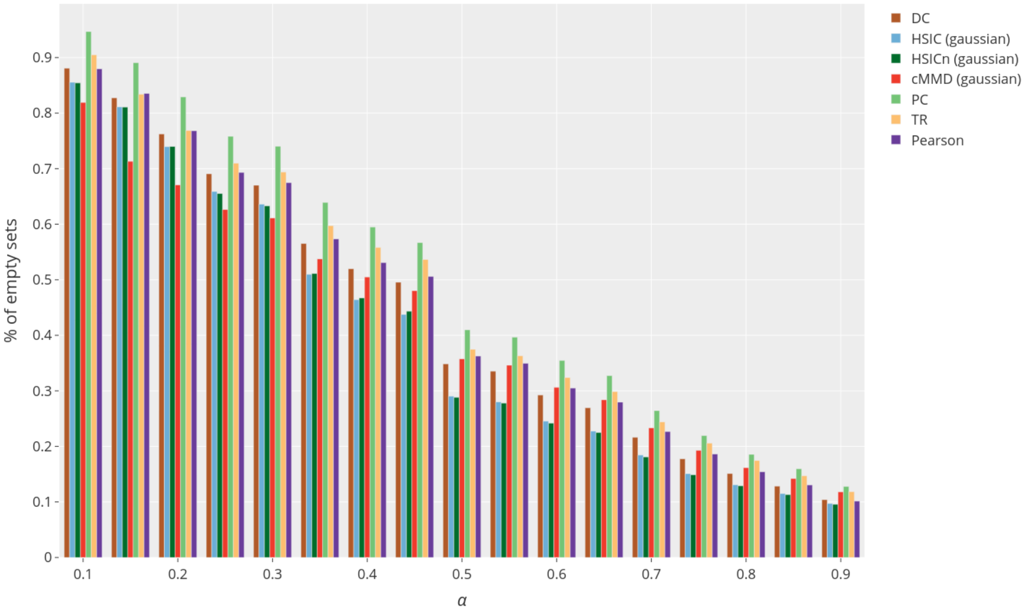

Figure 29: Percentages of empty sets by association measure for each α and pooled from the 200 batches of each DGPs.

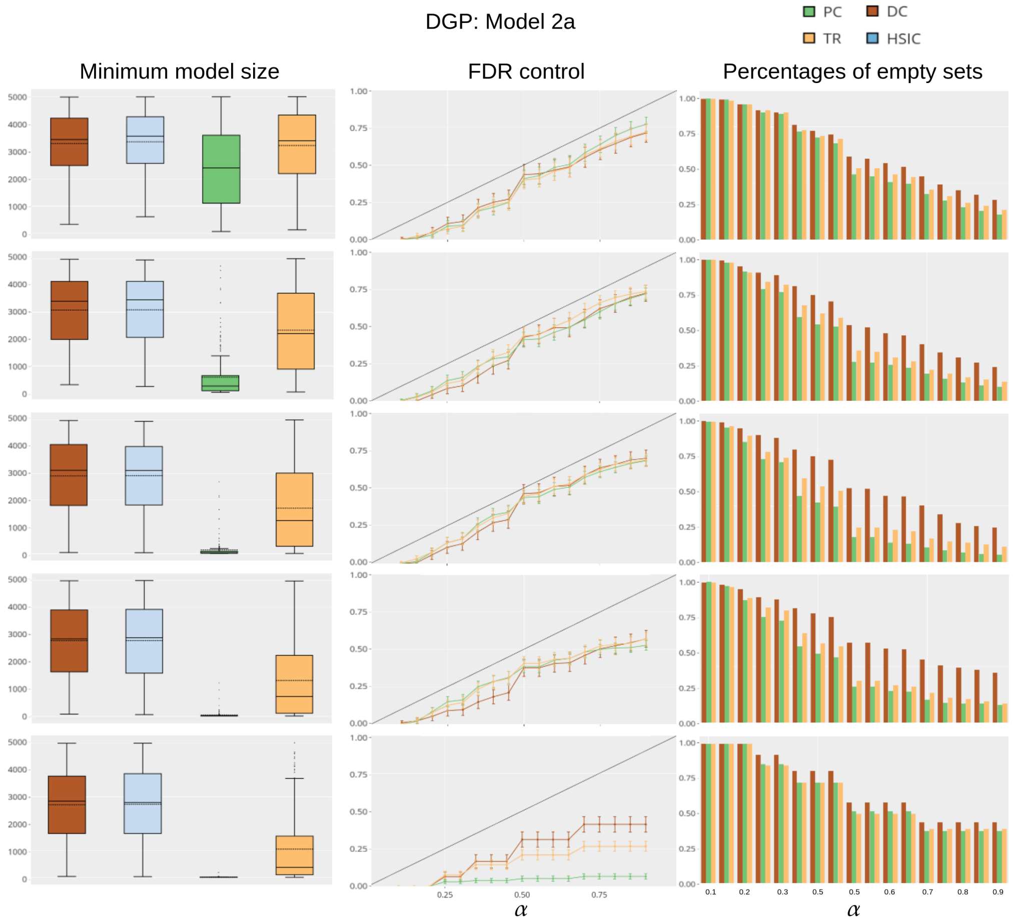

Figure 30: Minimum model size, FDR control and percentages of empty sets for DGP 2.a with $n = 100$ and $p = 5000$. Each row have respectively a value of n_0 equal to 0.1, 0.3, 0.5, 0.7 and 0.9 of n.

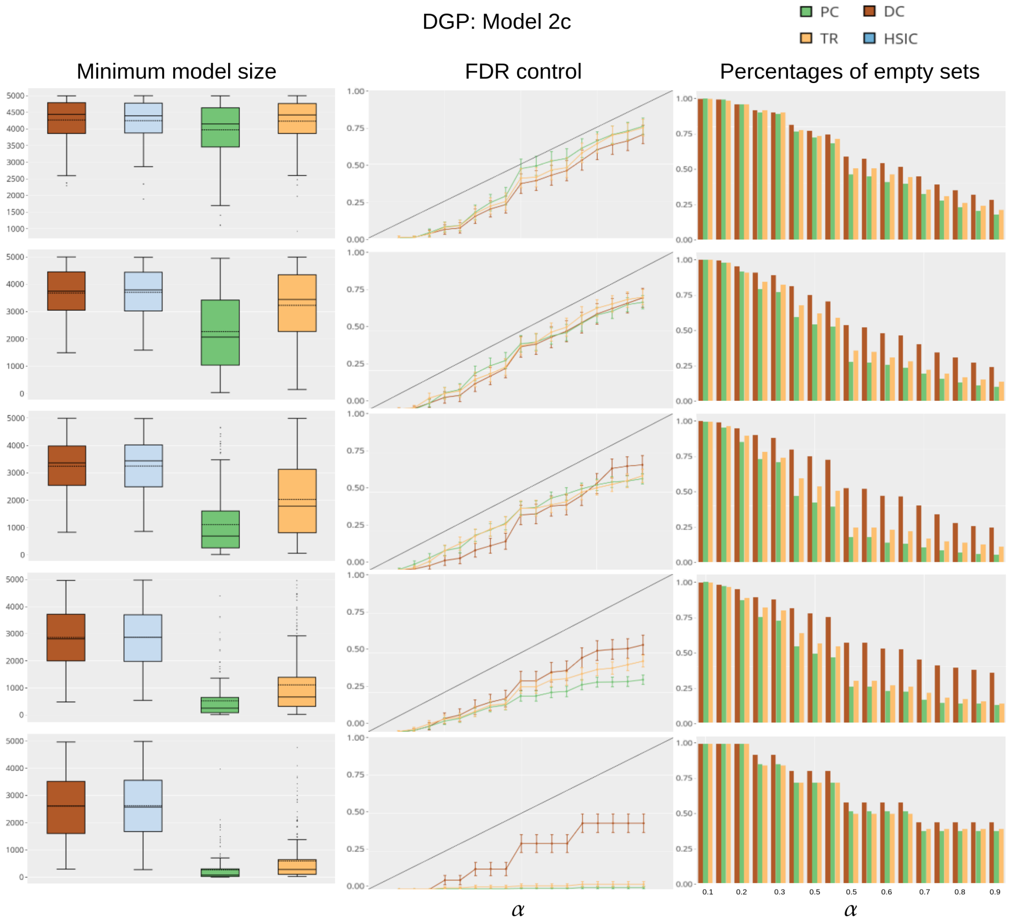

Figure 31: Minimum model size, FDR control and percentages of empty sets for DGP 2.c with $n = 100$ and $p = 5000$. Each row have respectively a value of n_0 equal to 0.1, 0.3, 0.5, 0.7 and 0.9 of n.

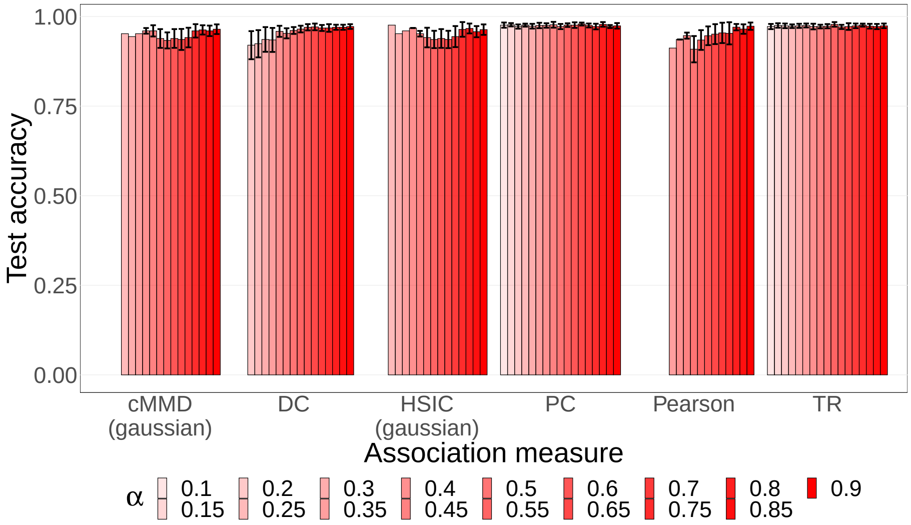

Figure 32: Test classification accuracy of a random forest trained exclusively on the pixels selected in the MNIST dataset. For each method and α, the protocol was run on 8 separate train-test splits. The following hyper-parameters of the random forest were set by cross-validation: number of trees (200, 500), the measure of quality of the splits (Gini impurity, entropy), the maximum depth of the trees (4, 6, 8) and the number of features to consider ($\log_2 p$, $\sqrt{p}$).

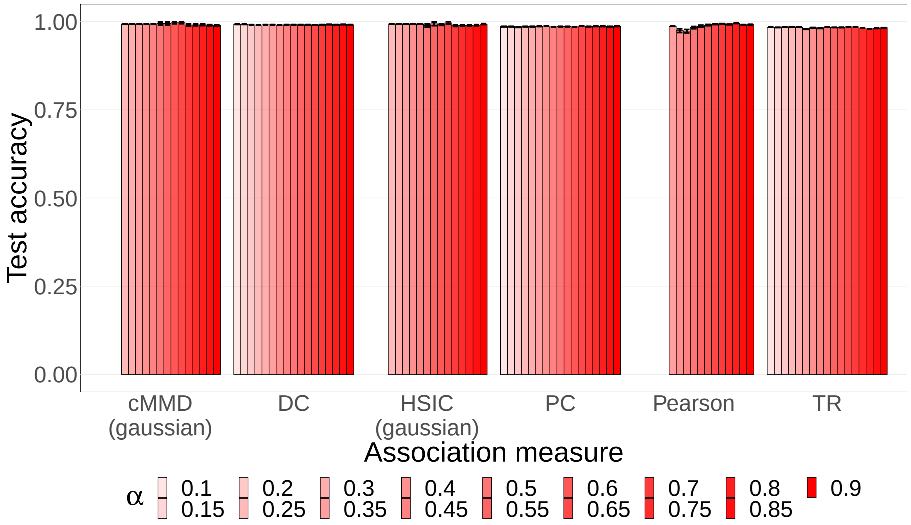

Figure 33: Test classification accuracy of a random forest trained exclusively on the genes selected in the TCGA BRCA dataset. For each method and α, the protocol was run on 8 separate train-test splits. The following hyper-parameters of the random forest were set by cross-validation: number of trees (200, 500), the measure of quality of the splits (Gini impurity, entropy), the maximum depth of the trees (4, 6, 8) and the number of features to consider ($\log_2 p$, $\sqrt{p}$).

E Breast cancer biomarkers

We ran the proposed knockoff procedure using three association measures (HSIC, cMMD and TR) using an $\alpha = 0.1$. The genes selected by any of them are available on Table 2.

In total, 261 unique genes were selected. Out of those, 55 (21%) are known cancer genes registered at the NCG 7.0 database (Dressler et al., 2021). Since only 3 337 out of 18 868 genes are included in the database, this overlap is high (Fisher's exact test one-sided P-value = 0.07). This was particularly true in the case of cMMD: out of 71 selected genes, 20 of them (28%) were known cancer genes (Fisher's exact test one-sided P-value = 0.02). This suggests that the knockoff procedure, using cMMD as association measure, can bring novel insights to biomarker discovery.

To gain insights into the broader mechanisms at play among the selected genes, we conducted a pathway enrichment analysis (Table 3). Only the TR measure and the joint analysis of all the selected genes resulted in significant pathways.

Table 2: 261 genes selected by the proposed pipeline, using cMMD, HSIC and TR as association measures.

AADACL2
AADACL3
ABCA10
ABCA8
ABCA9
ACE2
ACSM2A
ACSM2B
ADAM33
ADAMTS5
ADCYAP1R1
ADGRA1
ADGRG4
ADH1A
ADH1C
ADH4
ADH7
ADRB2
AKAIN1
AMOTL1
ANGPTL1
ANKRD29
ANKRD33
ANXA1
APBA1
ARF1
ARHGAP20
ARHGEF28
ASPA
BMX
BTNL3
BTNL9
C16orf82
C1orf185
C1QTNF9
C8orf88
CA4
CAB39L
CACHD1
CALHM4
CAPN11
CARD18
CAV1
CAV2
CAVIN2
CCDC178
CCL14
CCT3
CD300LG
CES1
CHRDL1
CLDN19
CNTNAP3B
COL17A1
CORO2B
CPA1
CSN1S1
CSN2
CYYR1
DAP3
DCAF12L1
DEFB118
DEFB119
DMD
DPRX
EDNRB
EGFR
ELAVL1
EOGT
EXOC1L
FABP9
FAM171A1
FAM181A
FAM184A
FAM205C
FAM236D
FAM71A
FEZF2
FGFR1
FLAD1
FMO2
FREM1
GABRA4
GFRAL
GKN1
GMNC
GOLGA8F
GOLGA8M
GPR149
GPR50
GPRC5B
GRIA4
GRXCR2
HIF3A
HLF
HOXA4
HOXA5
HSPB2-C11orf52
IFNA8
IGSF10
IL33
INMT-MINDY4
ISM1
IZUMO3
JAM2
KCNA4
KCNH5
KCTD4
KCTD9
KIAA0408
KL
KLHL29
KLHL33
KLHL40
KRT25
KRTAP1-1
KRTAP13-4
LALBA
LEPR
LHCGR
LIFR
LMOD1
LRIG3
LRRC2
LRRC3B
LRRN4CL
MAB21L1
MAMDC2
MAS1
MASP1
MATN2
MAZ
MEIS2
MFSD4A
MICU3
MID1
MME
MRGPRX2
MYH11
MYOC
NACC1
NDEL1
NEK2
NFIB
NKAPL
NPAP1
NPY2R
NPY4R
NR0B1
NR3C2
NUAK2
NUP210
NXNL1
OR10A3
OR10A6
OR2L3
OR4D6
OR5AP2
OR5P2
OSTN
OTC
OXTR
PAFAH1B3
PAK3
PALMD
PAMR1
PARP1
PCDH11Y
PCDHGB6
PDE1C
PDE2A
PEAR1
PELI2
PF4V1
PGA3
PGA4
PGA5
PHYHIP
PLA2G4A
PLD1
PLSCR4
PMP2
PPDPFL
PPM1F
PPP1R17
PPP4C
PRAMEF18
PRCC
PRKD1
PRKN
PROS1
PRRG3
PSG7
PTBP1
PYGO2
RABIF
RAX2
RBFOX2
RCBTB2
RGMA
RGN
RNF145
RUSC1
RYR3
SAV1
SCAMP3
SCN4B
SEC14L5
SEMA6A
SERPINB12
SERPINB13
SERTM1
SGCE
SGCZ
SH3BGRL2
SIRPA
SLC17A7
SLC22A12
SLC35A2
SLC50A1
SMIM21
SOBP
SPC25
SPRR2B
SPRR2F
SPRY2
SRPX
STOML3
SULT1C3
SVEP1
SYNM
SYNPO2
TANGO6
TBL2
TCF7L1
TGFBR2
TGFBR3
TIMM17A
TINAGL1
TMEFF2
TMEM220
TMEM252
TOR3A
TPM3
TRIM11
TRPM3
TSHB
TSHZ2
TSLP
TSPAN19
TTC23
UGT1A7
VAMP2
VEGFD
WASF2
XAGE3
XAGE5
YBX3
ZNF705G
ZSCAN4
ZSWIM2

Table 3: Pathways with an adjusted P-value < 0.05 in a pathway enrichment analyses on the selected genes by each of the methods. The studied pathways were the canonical pathways from the sets obtained from MSigDB's curated gene sets (v6.0). The gene universe was set to all protein coding genes. When the measure reads "Aggregated" means that we used the genes selected by any of the methods.

Description	Measure	pvalue	p.adjust	qvalue
WP FATTY ACID OMEGA OXIDATION	TR	3.7e-05	0.0052	0.0044
WP FATTY ACID OMEGA OXIDATION	Aggregated	4.1e-05	0.0353	0.0341
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	Aggregated	9.0e-05	0.0386	0.0373
KEGG RETINOL METABOLISM	TR	1.8e-04	0.0074	0.0063
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	TR	1.9e-04	0.0074	0.0063
KEGG METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	TR	2.5e-04	0.0074	0.0063
KEGG DRUG METABOLISM CYTOCHROME P450	TR	2.8e-04	0.0074	0.0063
REACTOME SURFACTANT METABOLISM	TR	3.1e-04	0.0074	0.0063
REACTOME BIOLOGICAL OXIDATIONS	TR	3.8e-04	0.0077	0.0065
KEGG FATTY ACID METABOLISM	TR	8.5e-04	0.0134	0.0113
KEGG TYROSINE METABOLISM	TR	8.5e-04	0.0134	0.0113
KEGG GLYCOLYSIS GLUCONEOGENESIS	TR	2.5e-03	0.0347	0.0293
REACTOME KERATINIZATION	TR	2.7e-03	0.0347	0.0293
REACTOME RA BIOSYNTHESIS PATHWAY	TR	4.3e-03	0.0475	0.0401
WP PEPTIDE GPCRS	TR	4.3e-03	0.0475	0.0401