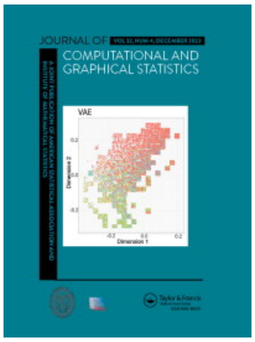

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Scalable Model-Free Feature Screening via Sliced-Wasserstein Dependency

Tao Li, Jun Yu & Cheng Meng

To cite this article: Tao Li, Jun Yu & Cheng Meng (2023) Scalable Model-Free Feature Screening via Sliced-Wasserstein Dependency, Journal of Computational and Graphical Statistics, 32:4, 1501-1511, DOI: 10.1080/10618600.2023.2183213

To link to this article: https://doi.org/10.1080/10618600.2023.2183213

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 32, NO. 4, 1501–1511
https://doi.org/10.1080/10618600.2023.2183213

Taylor & Francis
Taylor & Francis Group

Scalable Model-Free Feature Screening via Sliced-Wasserstein Dependency

Tao Li[a], Jun Yu[b], and Cheng Meng[c]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [b]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China; [c]Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China

ABSTRACT

We consider the model-free feature screening problem that aims to discard non-informative features before downstream analysis. Most of the existing feature screening approaches have at least quadratic computational cost with respect to the sample size n, thus, may suffer from a huge computational burden when n is large. To alleviate the computational burden, we propose a scalable model-free sure independence screening approach. This approach is based on the so-called sliced-Wasserstein dependency, a novel metric that measures the dependence between two random variables. Specifically, we quantify the dependence between two random variables by measuring the sliced-Wasserstein distance between their joint distribution and the product of their marginal distributions. For a predictor matrix of size $n \times d$, the computational cost for the proposed algorithm is at the order of $O(n \log(n)d)$, even when the response variable is multivariate. Theoretically, we show the proposed method enjoys both sure screening and rank consistency properties under mild regularity conditions. Numerical studies on various synthetic and real-world datasets demonstrate the superior performance of the proposed method in comparison with mainstream competitors, requiring significantly less computational time. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2022
Accepted January 2023

KEYWORDS
Multivariate response model; Nonlinear model; Optimal transport; Sure screening; Variable selection

1. Introduction

During recent decades, the rapid development of science and technologies has enabled researchers to collect data with ultrahigh-dimensional features. Such ultrahigh-dimensional datasets are emerging in all fields of science and engineering, from academia to industry (Tibshirani et al. 2003; Fan and Ren 2006; Weinberger et al. 2009; Pang et al. 2018; Li et al. 2022). These datasets provide researchers with unprecedented opportunities for data-driven decision-making and knowledge discoveries. Nevertheless, traditional statistical and machine learning algorithms may face big challenges when analyzing ultrahigh-dimensional data. In particular, when the features contain redundant and noisy information, estimating their functional relationship with the response may become quite challenging due to the computational burden, memory cost, statistical accuracy, and algorithmic stability (Fan et al. 2009; Hall and Miller 2009; Lv and Liu 2014).

To overcome such challenges caused by ultrahigh-dimensionality, one commonly used technique is the sure independence screening (SIS) (Fan and Lv 2008). The SIS technique aims to screen out redundant features in linear models by ranking their marginal Pearson correlations. It is known that such a technique enjoys the so-called sure screening property, which states the selected features contain all the informative ones with probability approaching one. As a result, SIS has become a popular feature screening technique in ultrahigh-dimensional studies (Liu et al. 2015; Liu and Li 2020). Recently, such a technique has been further extended from simple linear regression to other problems, which include generalized linear model (Fan and Song 2010), multi-index semi-parametric model (Zhu et al. 2011), nonparametric model (Fan et al. 2011; Liu et al. 2014), quantile regression (He et al. 2013; Wu and Yin 2015), compressed sensing (Xue and Zou 2011), among others. In addition, such a technique has been extended to the model-free setting (Zhu et al. 2011; Li et al. 2012; Mai and Zou 2015; Liu et al. 2020). These methods are known to enjoy the sure screening property without specifying a regression model.

While the existing model-free feature screening methods have already shown extraordinary performance, they may suffer from huge computational costs in practice. Take the popular distance correlation screening (DC-SIS) approach as an example (Li et al. 2012), for a predictor matrix of the size $n \times d$, the computational cost of DC-SIS is at the order of $O(n^2d)$. Another example is the recently proposed PC-SIS method, whose computational cost is of the order $O(n^3d)$ (Liu et al. 2020; Xu et al. 2020). Such computational costs hinder the wide application of model-free feature screening methods on large-scale datasets such that both n and d are considerable.

To alleviate the computational burden, we develop a scalable model-free feature screening method. This approach is based on the so-called sliced-Wasserstein (SW) dependency, a novel metric that measures the dependence between two random variables. The idea of the SW dependency is motivated by the

CONTACT Cheng Meng chengmeng@ruc.edu.cn Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

notion of mutual information (Kullback 1997). Consider two random variables X and Y with the marginal distributions μ and ν, respectively. Let γ be the joint distribution of X and Y and $\mu \otimes \nu$ be the product of their marginal distributions. The mutual information quantifies the dependency between X and Y by measuring the Kullback–Leibler divergence between γ and $\mu \otimes \nu$. In this paper, instead of using the Kullback–Leibler divergence, we consider the SW dependency that measures the sliced-Wasserstein distance between γ and $\mu \otimes \nu$. The sliced-Wasserstein distance is known to have nice theoretical properties and is easy to calculate (Rabin et al. 2011; Nadjahi 2021); details will be provided later. The prototypical problem in optimal transport is to evaluate the Wasserstein distances between distributions. Among different variants of the Wasserstein distance, the sliced-Wasserstein distance has attracted wide attention in the machine learning community and has been successfully applied in many tasks, such as data classification(Kolouri et al. 2016; Carriere et al. 2017), generative models (Wu et al. 2019; Deshpande et al. 2019; Kolouri et al. 2018; Meng et al. 2019; Xu et al. 2020), and Bayesian inference (Nadjahi et al. 2020). Many variants of sliced-Wasserstein distance have been proposed to improve efficiency, including maximum SW distances (Deshpande et al. 2019), generalized SW distances (Kolouri et al. 2019), orthogonal SW distances (Rowland et al. 2019), distributional SW distances (Nguyen et al. 2020), and Hilbert curve projection distance (Li et al. 2022).

Using the SW dependency, we develop a model-free feature screening algorithm by ranking the SW dependency with respect to (w.r.t.) the response and each feature, respectively. The proposed algorithm has a computational cost of the order $O(n\log(n)d)$, no matter whether the response variable is univariate or not. Theoretically, we show the algorithm enjoys sure screening and rank consistency properties under mild regularity conditions. To evaluate the empirical performance and computational time of the proposed method, we compare it with several mainstream competitors through extensive synthetic and real-world datasets. The numerical experiments show the proposed method yields comparable results, requiring significantly less CPU time.

The remainder of this article is organized as follows. We start in Section 2 by introducing the essential background of optimal transport and transport dependency. In Section 3, we present the details of the proposed sliced-Wasserstein dependency. We then develop a feature screening approach using the sliced-Wasserstein dependency. Details of the algorithm and theoretical properties of this approach are provided in Section 4. We examine the performance of the proposed method through extensive simulation and two real data examples in Sections 5 and 6, respectively. Section 7 concludes the article, and the technical proofs are provided in the supplementary materials.

2. Preliminaries

2.1. Optimal Transport and Wasserstein Distance

Let $\mathcal{P}(\mathbb{R}^d)$ be the set of probability measures on $\mathbb{R}^d$ and $\mathcal{P}_p(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^p d\mu(x) < \infty\}$ be the set of probability measures on $\mathbb{R}^d$ with finite moment of order p. Kantorovich (1942) considered a family of the joint distribution of μ and ν, termed as the "coupling" π, such that two marginal distributions of π are equal to μ and ν, respectively. Let Π be the set of all such couplings, that is,

$$\Pi(\mu,\nu) = \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)\ s.t.\ \forall \text{ Borel set } A, B \subset \mathbb{R}^d,\\ \pi(A \times \mathbb{R}^d) = \mu(A),\ \pi(\mathbb{R}^d \times B) = \nu(B)\}.$$

Among all the couplings $\pi \in \Pi(\mu,\nu)$, of interest is to find the optimal one, defined by

$$\pi^* := \underset{\pi\in\Pi(\mu,\nu)}{\arg\inf} \int \|x-y\|^p d\pi(x,y),$$

where $\|\cdot\|$ is the Euclidean norm. Such a minimization problem is called the optimal transport problem. Treating two distributions μ and ν as two masses, each coupling π could be seen as a way of moving one distribution of mass to another. The optimal transport is moving one distribution of mass to another as efficiently as possible. It has been widely studied in mathematics, probability, and economics; see Peyré and Cuturi (2019) and Panaretos and Zemel (2019) for recent reviews. Closely related to the optimal transport problem is the Wasserstein distance. In particular, the p-Wasserstein distance of two probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$W_p(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^p d\pi(x,y)\right)^{1/p}.$$

Wasserstein distance measures the "transport cost" between two measures. As a result, it is also called the earth mover's distance in the literature (Levina and Bickel 2001; Peyré and Cuturi 2019).

One interesting fact about the Wasserstein distance is that it admits a closed-form expression for one-dimensional measures. Specifically, consider two one-dimensional measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, the p-Wasserstein distance ($p \geq 1$) between them takes the form

$$W_p(\mu,\nu) = \left(\int_0^1 \left|F_\mu^{-1}(x) - F_\nu^{-1}(x)\right|^p dx\right)^{1/p}, \tag{1}$$

where F_μ and F_ν are the cdf w.r.t. μ and ν, respectively. This closed-form expression indicates that one can calculate one-dimensional Wasserstein distances through sorting, requiring a $O(n\log(n))$ computational cost.

2.2. Sliced-Wasserstein Distance

Using the closed-form expression in (1), researchers developed the sliced-Wasserstein distance for high-dimensional cases (Rabin et al. 2011). Loosely speaking, the SW distance uses the random projection technique to break down the high-dimensional problem into a series of subproblems, each of which involves the calculation of a one-dimensional Wasserstein distance.

More formally, let $\mathbb{S}^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : ||\boldsymbol{u}|| = 1\}$ be the d-dimensional unit sphere, and let $\langle\cdot,\cdot\rangle$ be the Euclidean inner-product. For any $\boldsymbol{u} \in \mathbb{S}^{d-1}$, let $\boldsymbol{u}^*$ be the linear form w.r.t. $\boldsymbol{u}$, such that for $\boldsymbol{a} \in \mathbb{R}^d$, $\boldsymbol{u}^*(\boldsymbol{a}) = \langle \boldsymbol{u}, \boldsymbol{a}\rangle$. For any measurable function $\phi : \mathbb{R}^d \to \mathbb{R}$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\phi_\# \mu$ is the push-forward measure of μ by ϕ : for any Borel set Ω in $\mathbb{R}$, $\phi_\#\mu(\Omega) = \mu\left(\phi^{-1}(\Omega)\right)$, with $\phi^{-1}(\Omega) = \{x \in \mathbb{R}^d : \phi(x) \in \Omega\}$. For any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the SW distance w.r.t. L_p norm is defined as

$$SW_p(\mu, \nu) := \left(\int_{\mathbb{S}^{d-1}} W_p^p(\boldsymbol{u}^*_\# \mu, \boldsymbol{u}^*_\# \nu) d\sigma(\boldsymbol{u}) \right)^{1/p}, \qquad (2)$$

where σ represents the uniform distribution on $\mathbb{S}^{d-1}$.

In practice, the integration in (2) can be approximated using a Monte Carlo scheme, that is, one can randomly and uniformly draw a finite set of projection directions from $\mathbb{S}^{d-1}$, and replace the integral with a finite-sample average (Rabin et al. 2011; Bonneel et al. 2015). Algorithm 1 summarizes the empirical algorithm to estimate the sliced-Wasserstein distance.

Algorithm 1 Empirical p-sliced-Wasserstein distance estimation

```
**Input:** $\{\mathbf{x}_i\}_{i=1}^n \sim \mu,\ \{\mathbf{y}_i\}_{i=1}^n \sim \nu, L, p \geq 1$
**Initialize** $D \leftarrow 0$
**for** $l = 1 : L$ **do**
    (i) Generate a random vector $\boldsymbol{u}_l$ from $\mathbb{S}^{d-1}$
    (ii) Compute $\hat{\mathbf{x}}_i = \langle \boldsymbol{u}_l, \mathbf{x}_i \rangle$ and $\hat{\mathbf{y}}_i = \langle \boldsymbol{u}_l, \mathbf{y}_i \rangle$ for $i = 1, \dots, n$
    (iii) Sort $\{\hat{\mathbf{x}}_i\}_{i=1}^n$ and $\{\hat{\mathbf{y}}_i\}_{i=1}^n$ in ascending order, denoted by $\{\hat{\mathbf{x}}_{[i]}\}_{i=1}^n$ and $\{\hat{\mathbf{y}}_{[i]}\}_{i=1}^n$, respectively
    (iv) $D \leftarrow D + \sum_{i=1}^n (\hat{\mathbf{x}}_{[i]} - \hat{\mathbf{y}}_{[i]})^p / nL$
**end for**
**Output:** $D^{1/p}$
```

2.3. Transport Dependency

Consider the problem of how to measure the (nonlinear) dependence between two random variables. Let X and Y be two random variables with the marginal distributions μ and ν, respectively. Let γ be the joint distribution of X and Y and $\mu \otimes \nu$ be the product of their marginal distributions. Intuitively, one has $\gamma = \mu \otimes \nu$ if and only if X and Y are independent. On the other hand, one would expect γ to be significantly different from $\mu \otimes \nu$ if X and Y are strongly dependent. Following this line of thinking, a natural way to quantify the dependence between X and Y is to measure some kind of "distance" between γ and $\mu \otimes \nu$. One famous measurement of this kind is the mutual information, defined as $D_{KL}(\gamma | \mu \otimes \nu)$, where $D_{KL}(\cdot|\cdot)$ denotes that Kullback-Leibler divergence, that is, KL-divergence. Such a measurement has found broad application in independent component analysis (Stone 2004), feature selection (Peng et al. 2005), generative adversarial network (Belghazi et al. 2018), and representation learning (Bachman et al. 2019).

Recently, some literature suggested replacing the KL-divergence with the Wasserstein distance (Arjovsky et al. 2017; Ozair et al. 2019), resulting in the family of transport dependency. We illustrate the idea of transport dependency through a toy example in Figure 1. The idea of transport dependency was first proposed by Ozair et al. (2019) and was extended by different authors with different names and slightly different formulations, including Wasserstein correlation coefficients (WCC) (Wiesel 2022), Wasserstein dependence coefficients (WDC) (Mordant and Segers 2022), and transport correlations (TC) (Nies et al. 2021). We summarize these methods in Table 1, where the second column shows the explicit formulation of these methods. To get a better understanding of these formulations, one natural way is to consider two Gaussian variables such that their joint distribution is a two-dimensional Gaussian with correlation ρ. Under such a scenario, all the existing transport dependency methods have closed-form expressions (when $p = 2$), which are listed in the third column in Table 1. In addition, we illustrate these expressions in Figure 2 by comparing them with the classical Pearson correlation and the distance correlation (Székely et al. 2007). One can observe that all of these measurements show a similar pattern with the Pearson correlation and the distance correlation. Such an observation indicates the existing transport dependency methods can effectively measure the dependency between random variables.

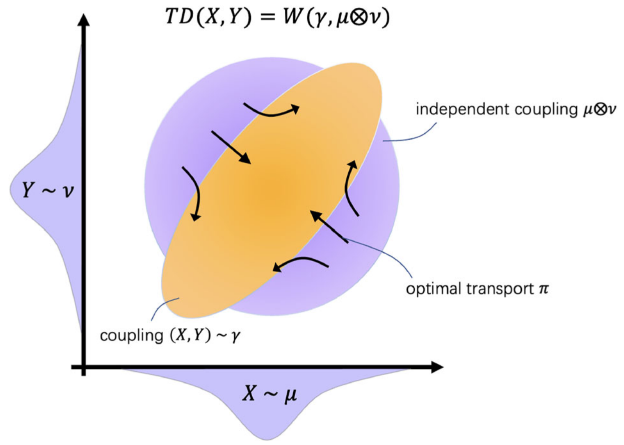
$TD(X,Y) = W(\gamma, \mu\otimes\nu)$
independent coupling $\mu\otimes\nu$
$Y \sim \nu$
optimal transport π
coupling $(X,Y) \sim \gamma$
$X \sim \mu$

Figure 1. Intuitively, the idea of transport dependency is to quantify the dependence between two random variables by measuring the Wasserstein distance between their joint distribution γ and the product of their marginal distribution $\mu \otimes \nu$.

Table 1. Different dependency measures based on Wasserstein distance.

Name	Formulation	2-D Gaussian with $p = 2$	Complexity	References
WCC$_p$	$\left(\frac{\int W_p^p(\gamma_x, \nu)\mu(dx)}{\int \lVert x_1 - x_2 \rVert^p \mu(dx_1)\mu(dx_2)} \right)^{1/p}$	$\left(1 - \sqrt{1-\rho^2}\right)^{1/2}$	$n^3 \log(n)$	Wiesel (2022)
WDC$_p$	$\frac{W_p(\gamma, \mu\otimes\nu)}{\sup_{\tilde{\gamma}\in\Pi(\mu,\nu)} W_p(\tilde{\gamma}, \mu\otimes\nu)}$	$\left(\frac{2-\sqrt{2+2\sqrt{1-\rho^2}}}{2-\sqrt{2}} \right)^{1/2}$	/	Mordant and Segers (2022)
TC$_p$	$\frac{W_p(\gamma, \mu\otimes\nu)}{(\int \lVert x_1 - x_2 \rVert^p \mu(dx_1)\mu(dx_2))^{1/p}}$	$\left(2 - \sqrt{2 + 2\sqrt{1-\rho^2}}\right)^{1/2}$	$n^3 \log(n)$	Nies et al. (2021)
SWC$_p$	$\frac{SW_p(\gamma, \mu\otimes\nu)}{\sqrt{SW_p(\gamma_X, \mu\otimes\mu) SW_p(\gamma_Y, \nu\otimes\nu)}}$	$\left(\frac{\pi - \int_0^\pi \sqrt{1+\rho\sin(2\theta)}d\theta}{\pi - 2\sqrt{2}} \right)^{1/2}$	$n \log(n)$	proposed method

In this table, γ_x is the distribution of Y condition on $X = x$, γ_X is the joint distribution of (X, X), γ_Y is the joint distribution of (Y, Y), and ρ is the correlation of 2-D Gaussian.

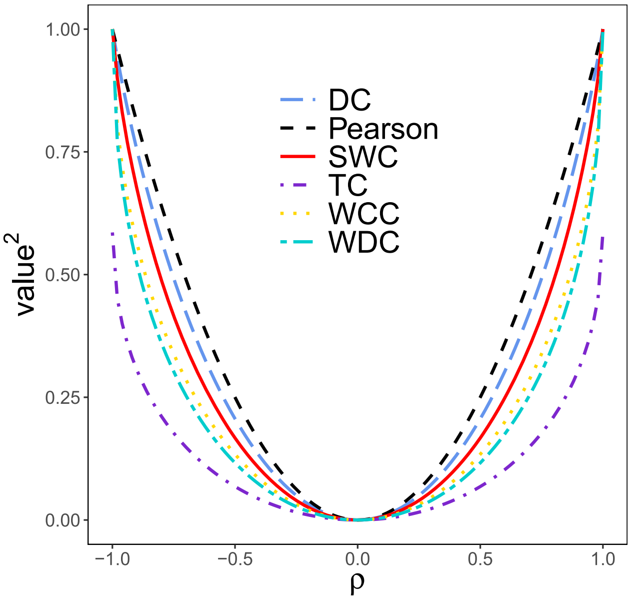

Figure 2. Comparison of different dependency measurements with $p = 2$ for two Gaussian variables, whose joint distribution is a two-dimensional Gaussian with correlation ρ. DC: distance correlation. Pearson: Pearson correlation. SWC: sliced-Wasserstein correlation(proposed). TC: transport correlation. WCC: Wasserstein correlation coefficients. WDC: Wasserstein dependence coefficient.

Despite the effectiveness, one limitation of the existing transport dependency methods is the huge computational burden. In particular, all of the three existing methods listed in Table 1 require calculating the Wasserstein distance, which is known to have the computational cost of the order $O(n^3 \log(n))$ for a sample of size n (Peyré and Cuturi 2019). In addition, the WDC approach involves a nontrivial optimization problem in its formulation, resulting in extra computational time. Although in practice, Wasserstein distance could be replaced by Sinkhorn divergence for faster computation (Genevay et al. 2019; Chizat et al. 2020), it is not known whether the theoretical properties of the transport dependency can still be preserved. Efficient and effective transport dependency measurement is still meager.

3. Sliced-Wasserstein Dependency

In this article, we develop an efficient dependency measurement called sliced-Wasserstein dependency. Let $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ be two random vectors with the marginal distributions μ and ν, respectively. Let γ be the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ and $\mu \otimes \nu$ be the product of their marginal distributions. The sliced-Wasserstein dependency (SWD) is defined as

$$\mathrm{SWD}_p(\mathbf{x}, \mathbf{y}) := SW_p(\gamma, \mu \otimes \nu),$$

where $SW_p(\gamma, \mu \otimes \nu)$ is the sliced p-Wasserstein distance between γ and $\mu \otimes \nu$. Let γ_μ and γ_ν be the joint distribution of $(\mathbf{x}, \mathbf{x})$ and $(\mathbf{y}, \mathbf{y})$, respectively. We also consider a normalized version of the sliced-Wasserstein dependency, called sliced-Wasserstein correlation (SWC), defined as

$$\mathrm{SWC}_p(\mathbf{x}, \mathbf{y}) := \frac{\mathrm{SWD}_p(\mathbf{x}, \mathbf{y})}{\sqrt{\mathrm{SWD}_p(\mathbf{x}, \mathbf{x})\mathrm{SWD}_p(\mathbf{y}, \mathbf{y})}} = \frac{SW_p(\gamma, \mu \otimes \nu)}{\sqrt{SW_p(\gamma_\mu, \mu \otimes \mu) SW_p(\gamma_\nu, \nu \otimes \nu)}},$$

where we follow the convention $0/0 = 0$.

We develop an empirical algorithm to estimate the sliced-Wasserstein correlation in practice. Let $\mathcal{I}_{\text{full}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{2n}$ be a sample generated from the joint distribution γ, where $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{2n}$ and $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^{2n}$ are generated from the distribution μ and ν, respectively. The key idea behind the proposed algorithm is that one can construct a sample that follows the distribution $\mu \otimes \nu$ using the observed sample $\mathcal{I}_{\text{full}}$ (Dai et al. 2022). To achieve the goal, we randomly split $\mathcal{I}_{\text{full}}$ into two equal-size sub-samples, denoted by $\mathcal{I} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ and $\tilde{\mathcal{I}} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{n}$. Further, we introduce the following notations

$$\begin{aligned} \mathcal{I}_{\mathbf{xy}} &= \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}, & \tilde{\mathcal{I}}_{\mathbf{xy}} &= \{(\tilde{\mathbf{x}}_i, \mathbf{y}_i)\}_{i=1}^{n}, \\ \mathcal{I}_{\mathbf{xx}} &= \{(\mathbf{x}_i, \mathbf{x}_i)\}_{i=1}^{n}, & \tilde{\mathcal{I}}_{\mathbf{xx}} &= \{(\tilde{\mathbf{x}}_i, \mathbf{x}_i)\}_{i=1}^{n}, \\ \mathcal{I}_{\mathbf{yy}} &= \{(\mathbf{y}_i, \mathbf{y}_i)\}_{i=1}^{n}, & \tilde{\mathcal{I}}_{\mathbf{yy}} &= \{(\tilde{\mathbf{y}}_i, \mathbf{y}_i)\}_{i=1}^{n}. \end{aligned}$$

It is obvious that the sample $\tilde{\mathcal{I}}_{\mathbf{xy}} = \{(\tilde{\mathbf{x}}_i, \mathbf{y}_i)\}_{i=1}^{n}, \tilde{\mathcal{I}}_{\mathbf{xx}} = \{(\tilde{\mathbf{x}}_i, \mathbf{x}_i)\}_{i=1}^{n}, \tilde{\mathcal{I}}_{\mathbf{yy}} = \{(\tilde{\mathbf{y}}_i, \mathbf{y}_i)\}_{i=1}^{n}$ follows the distribution $\mu \otimes \nu$, $\mu \otimes \mu$, $\nu \otimes \nu$, respectively. Let $I_{\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}}$ be the discrete measure that assigns mass $1/n$ to each observation $(\mathbf{x}_i, \mathbf{y}_i)$. The empirical sliced-Wasserstein dependency and sliced-Wasserstein correlation thus can be calculated as

$$\begin{aligned} \widehat{\mathrm{SWD}_p}(\mathbf{x}, \mathbf{y}) &:= SW_p(I_{\mathcal{I}_{\mathbf{xy}}}, I_{\tilde{\mathcal{I}}_{\mathbf{xy}}}), \\ \widehat{\mathrm{SWC}_p}(\mathbf{x}, \mathbf{y}) &:= \frac{SW_p(I_{\mathcal{I}_{\mathbf{xy}}}, I_{\tilde{\mathcal{I}}_{\mathbf{xy}}})}{\sqrt{SW_p(I_{\mathcal{I}_{\mathbf{xx}}}, I_{\tilde{\mathcal{I}}_{\mathbf{xx}}}) SW_p(I_{\mathcal{I}_{\mathbf{yy}}}, I_{\tilde{\mathcal{I}}_{\mathbf{yy}}})}}. \end{aligned}$$

We summarize a few nice properties of sliced-Wasserstein dependence and sliced-Wasserstein correlation as follows. The technical proof is relegated to supplementary material.

Theorem 1. For two random vectors $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$, we have

1. $\mathrm{SWD}_p(\mathbf{x}, \mathbf{y}) \geq 0$, the equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent. In addition, $\mathrm{SWD}_p(\mathbf{x}, \mathbf{x}) = 0$ implies that $\mathbf{x} = \mathbb{E}(\mathbf{x})$ almost surely.
2. $\mathrm{SWC}_p(\mathbf{x}, \mathbf{y}) \geq 0$, and equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent.
3. If $\mathbf{y} = \mathbf{a} + \mathbf{x}\mathbf{C}$, then $\mathrm{SWC}_p(\mathbf{x}, \mathbf{y}) = 1$ for all constant vectors $\mathbf{a}$ in $\mathbb{R}^{d_x}$, and $d_x \times d_x$ orthonormal matrices $\mathbf{C}$.
4. $\mathrm{SWC}_p(\mathbf{a_1} + b\mathbf{x}\mathbf{C_1}, \mathbf{a_2} + b\mathbf{y}\mathbf{C_2}) = \mathrm{SWC}_p(\mathbf{x}, \mathbf{y})$ for all constant vectors $\mathbf{a_1}$ in $\mathbb{R}^{d_x}$, $\mathbf{a_2}$ in $\mathbb{R}^{d_y}$, scalars b, $d_x \times d_x$ orthonormal matrices $\mathbf{C_1}$, and $d_y \times d_y$ orthonormal matrices $\mathbf{C_2}$.

Theorem 2. If there exists a constant $\delta > 2$ s.t. $\mathbb{E}\|\mathbf{x}\|_{p\delta}^{p\delta} < \infty$ and $\mathbb{E}\|\mathbf{y}\|_{p\delta}^{p\delta} < \infty$, then almost surely

$$\begin{aligned} \lim_{n \to \infty} \widehat{\mathrm{SWD}_p}(\mathbf{x}, \mathbf{y}) &= \mathrm{SWD}_p(\mathbf{x}, \mathbf{y}), \\ \lim_{n \to \infty} \widehat{\mathrm{SWC}_p}(\mathbf{x}, \mathbf{y}) &= \mathrm{SWC}_p(\mathbf{x}, \mathbf{y}), \end{aligned}$$

where $\| \cdot \|_p$ is the p-norm.

We provide the following theorem to show its relationship with Pearson correlation in the bivariate normal distribution case.

Theorem 3. If X and Y are standard normal random variables with $\mathrm{cor}(X, Y) = \rho$, then

1. $\mathrm{SWC}_p^p(X, Y) = \frac{\int_0^{2\pi} |\sqrt{1+\rho\sin(2\theta)}-1|^p d\theta}{\int_0^{2\pi} |\sqrt{1+\sin(2\theta)}-1|^p d\theta}$.
2. $\mathrm{SWC}_p(X, Y)$ is a strictly monotonically increasing function of $|\rho|$.
3. $\mathrm{SWC}_p(X, Y) \leq |\rho|$ and $\lim_{\rho\to 0} \frac{\mathrm{SWC}_p(X,Y)}{|\rho|} = \left(\frac{\int_0^{2\pi} |\sin(2\theta)|^p d\theta}{\int_0^{2\pi} |2\sqrt{1+\sin(2\theta)}-2|^p d\theta}\right)^{1/p}$.

Remark 1. Specifically, if $p = 2$, we have

$$\mathrm{SWC}_2^2(X, Y) = \frac{\pi - \int_0^{\pi} \sqrt{1+\rho\sin(2\theta)}d\theta}{\pi - 2\sqrt{2}},$$

$$\lim_{\rho\to 0} \frac{\mathrm{SWC}_2(X, Y)}{|\rho|} = \frac{\sqrt{\pi}}{4\sqrt{\pi - 2\sqrt{2}}} \approx 0.79182.$$

The following theorem demonstrates the asymptotic properties of sliced-Wasserstein dependence when $p = 1$. We leave the case $p > 1$ as future research. We may use this theorem to conduct an independence test.

Theorem 4. Let $\mathbf{x}, \mathbf{y}$ be two random vectors. Let $\tilde{\mathbf{x}}$ be a random vector that has the same distribution as $\mathbf{x}$ and is independent of $\mathbf{x}$ and $\mathbf{y}$. Denote $\mathbf{z}_1 = (\mathbf{x}, \mathbf{y})$, $\mathbf{z}_2 = (\tilde{\mathbf{x}}, \mathbf{y})$. Let $F(\boldsymbol{\theta}, t) = \Pr(\mathbf{z}_1\boldsymbol{\theta}^T \leq t)$, $G(\boldsymbol{\theta}, t) = \Pr(\mathbf{z}_2\boldsymbol{\theta}^T \leq t)$ and G' is a centered Gaussian process with a covariance function

$$\begin{aligned}&\mathrm{cov}\left(G'(\boldsymbol{\theta}_1, t_1), G'(\boldsymbol{\theta}_2, t_2)\right)\\&\quad = \mathrm{cov}\left(\mathbf{1}_{\{\mathbf{z}_1\boldsymbol{\theta}_1^T \leq t_1\}} - \mathbf{1}_{\{\mathbf{z}_2\boldsymbol{\theta}_1^T \leq t_1\}}, \mathbf{1}_{\{\mathbf{z}_1\boldsymbol{\theta}_2^T \leq t_2\}} - \mathbf{1}_{\{\mathbf{z}_2\boldsymbol{\theta}_2^T \leq t_2\}}\right).\end{aligned}$$

where $\mathbf{1}_A$ is the indicator function of set A. Assume that there exists a constant $\delta > 0$ s.t. $\mathbb{E}\|\mathbf{x}\|_{2+\delta}^{2+\delta} < \infty$ and $\mathbb{E}\|\mathbf{y}\|_{2+\delta}^{2+\delta} < \infty$. Then, we have

(i) if $\mathbf{x}$ and $\mathbf{y}$ are independent, then

$$\sqrt{n}\widehat{\mathrm{SWD}_1}(\mathbf{x}, \mathbf{y}) \xrightarrow{d} \int |G'| \, dt d\sigma(\boldsymbol{\theta}).$$

(ii) if $\mathbf{x}$ and $\mathbf{y}$ are dependent, then

$$\sqrt{n}\widehat{\mathrm{SWD}_1}(\mathbf{x}, \mathbf{y}) \xrightarrow{\text{a.e.}} \infty.$$

4. Scalable Model-Free Feature Screening

We develop two model-free feature screening approaches based on the proposed Sliced-Wasserstein Dependency (SWD) and the Sliced-Wasserstein Correlation (SWC), respectively. These two approaches are completely model-free as they allow for arbitrary regression relationship of response onto the features, regardless of whether it is linear or nonlinear. They also permit univariate and multivariate responses, regardless of whether it is continuous, discrete, or categorical. Due to the space limitation, we mainly focus on the SWD-based approach in this section by introducing the algorithm and its theoretical properties. Details for the SWC-based approach are relegated to supplementary material.

4.1. Feature Screening with Sliced-Wasserstein Dependency

Let $\mathbf{y} = (Y_1, \ldots, Y_{d_y})^\intercal$ be a d_y-dimensional response vector and $\mathbf{x} = (X_1, \ldots, X_d)^\intercal$ be a d-dimensional vector of features. We focus on the scenario that d_y is fixed and $d \gg n$. Naturally, we may assume that only a small portion of the features are relevant to the response. Let $F(\mathbf{y}|\mathbf{x})$ be the conditional distribution function of $\mathbf{y}$ given $\mathbf{x}$. Following the notation in the literature (Li et al. 2012; Liu et al. 2020), without specifying any regression model of $\mathbf{y}$ given $\mathbf{x}$, we define the index set of the active features by

$$\mathcal{A} = \{k : F(\mathbf{y}|\mathbf{x}) \text{ functionally depends on } X_k, k = 1, \ldots, d\}.$$

Such a setting abstracts a large number of sparse regression problems, including linear models and nonlinear models, among others. What's more, the multivariate response and grouped features are also allowed. Here, we denote the complement of $\mathcal{A}$ as $\mathcal{A}^c$.

Suppose we observe a random sample of size $2n$ from $(\mathbf{x}, \mathbf{y})$ and randomly split it into two halves. Without loss of generality, we denote the two sub-samples as $\mathcal{I} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ and $\tilde{\mathcal{I}} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^n$. For $k = 1, \ldots, d$, let X_{ik} and $\tilde{X}_{ik}$ be the kth column of $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$, respectively. Further, we introduce the following notations

$$\mathcal{I}_{X_k\mathbf{y}} = \{(X_{ik}, \mathbf{y}_i)\}_{i=1}^n, \quad \tilde{\mathcal{I}}_{X_k\mathbf{y}} = \{(\tilde{X}_{ik}, \mathbf{y}_i)\}_{i=1}^n.$$

Recall that X_k and $\mathbf{y}$ are independent if and only if the sliced-Wasserstein dependency $\mathrm{SWD}_1(X_k, \mathbf{y})$ equals zero. This motivates us to screen out the features X_i such that the value $\mathrm{SWD}_1(X_i, \mathbf{y})$ is relatively small. To be specific, we compute the empirical sliced-Wasserstein dependency between X_k and $\mathbf{y}$ as

$$\widehat{\mathrm{SWD}_1}(X_k, \mathbf{y}) = SW_1(I_{\mathcal{I}_{X_k\mathbf{y}}}, I_{\tilde{\mathcal{I}}_{X_k\mathbf{y}}}).$$

Then, we propose to estimate the active set $\mathcal{A}$ by

$$\widehat{\mathcal{A}}_1 = \{k : \widehat{\mathrm{SWD}_1}(X_k, \mathbf{y}) \geq c_1 n^{-c_2}, \ 1 \leq k \leq d\},$$

where c_1 and c_2 are prespecified threshold values, which will be defined in Condition B.1 later. We name this approach as sliced-Wasserstein dependency screening, summarized in Algorithm 2.

Algorithm 2 Sliced-Wasserstein dependency screening

```
Input: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{2n}, c_1, c_2$
for $k = 1 : d$ do
    Calculate $\widehat{\mathrm{SWD}_1}(X_k, \mathbf{y})$ using Algorithm 1
end for
$\widehat{\mathcal{A}}_1 = \{k : \widehat{\mathrm{SWD}_1}(X_k, \mathbf{y}) \geq c_1 n^{-c_2}, \ 1 \leq k \leq d\}$
Output: $\widehat{\mathcal{A}}_1$
```

4.2. Theoretical Results

We now study the theoretical properties of SWD screening. In this section, we take the sliced-Wasserstein dependency in Algorithm 1 as the true value. Under some regularity conditions, we show this method enjoys the sure screening property (see Theorem 5), which states that with probability approaching one, all active features are included in $\widehat{\mathcal{A}}_1$. We also show that it has the rank consistency property (see Theorem 6), which states that almost surely the active features are ranked ahead of the inactive ones as the sample size tends to infinity. We provide two different conditions (see Conditions A.1 and A.2) which allow for different orders of dimension d.

For $\alpha > 0$, we define $\|\xi\|_{\psi_\alpha} := \inf\{C > 0 : E[\exp((|\xi|/C)^\alpha)] \leq 2\}$ for a real-valued random variable ξ. The following regularity conditions help accomplish the proof, which may not be the weakest ones.

Condition A.

(A1). Assume that it holds uniformly for d that $\max_{1\leq j\leq d}\|X_j\|_{\psi_\alpha} < M$ and $\|\|\mathbf{y}\|_1\|_{\psi_\alpha} < M$ for some $\alpha \in (0,1]$ and $M > 0$.

(A2). Assume that it holds uniformly for d that $\max_{1\leq j\leq d}\mathbb{E}|X_j|^q < M$ and $\mathbb{E}\|\mathbf{y}\|_q^q < M$ for some $q \in (2,\infty)$ and $M > 0$.

Condition A.1 is weaker than Condition (C1) in Li et al. (2012). It has weaker restrictions on the tail distribution. Indeed, Condition (C1) in Li et al. (2012) clarifies that statements in Condition A.1 is satisfied when $\alpha = 2$, which can be easily induced when $\alpha \in (0,1]$. Condition A.1 follows immediately when X_j and $\mathbf{y}$ are sub-Gaussian random variables or subexponential random variables. Condition A.2 is even weaker than Condition A.1 which only needs the existence of q-order moment ($q > 2$). As a result, theoretical results based on Condition A.2 allow for a lower order of d.

Condition B.

(B1). The minimum SWD_1 of active features satisfies

$$\min_{k\in\mathcal{A}}\mathrm{SWD}_1(X_k,\mathbf{y}) \geq 2c_1 n^{-c_2}$$

for some constants $c_1 > 0$ and $0 \leq c_2 < 1/2$.

(B2).

$$\liminf_{d\to\infty}\left\{\min_{k\in\mathcal{A}}\mathrm{SWD}_1(X_k,\mathbf{y}) - \max_{k\in\mathcal{A}^c}\mathrm{SWD}_1(X_k,\mathbf{y})\right\} \geq 2c_3$$

where $c_3 > 0$ is a constant.

Condition B.1 is quite common in the marginal screening literature, and it is the same as the Condition 3 in Fan and Lv (2008) and the Condition(C2) in Li et al. (2012). This condition guarantees that the dependency between the active features and response cannot converge to zero too fast as n diverges. Condition B.2 is also quite mild, and it is similar to Condition 3 in Cui et al. (2015). This condition imposes that there exists a gap of signal between active and inactive features.

Theorem 5 gives sure screening property under two different conditions. Condition A.1 allows for exponential order of d, while less restrictive Condition A.2 allows for polynomial order of d.

Theorem 5 (Sure screening).

(I) Assume Conditions A.1 and B.1 are satisfied. And if there exists a constant $\epsilon > 0$ such that $\log(d) = o\left(\frac{n^{\min\{1-2c_2-2\epsilon,\alpha(1-c_2-\epsilon)\}}}{\log(1+n)}\right)$, then we have as $n\to\infty$

$$\Pr(\mathcal{A}\subset\widehat{\mathcal{A}}_1)\to 1$$

(II) Assume Conditions A.2 and B.1 are satisfied. And if there exists a constant $\epsilon > 0$ such that $d = o(n^{q-1-qc_2-q\epsilon})$, then we have as $n\to\infty$

$$\Pr(\mathcal{A}\subset\widehat{\mathcal{A}}_1)\to 1$$

Theorem 6 provides a stronger theoretical result than sure screening property. Though it requires a more restrictive condition on the difference between active and inactive signals, it tells us that, almost surely, the active features are ranked ahead of the inactive ones as the sample size diverges.

Theorem 6 (Rank consistency).

(I) Assume Conditions A.1 and B.2 are satisfied. If $\log d = o\left(n^{\min\{1-2\epsilon_0,\alpha(1-\epsilon_0)\}}\right)$ for some constant $\epsilon_0 > 0$, then we have almost surely

$$\liminf_{n\to\infty}\left(\min_{k\in\mathcal{A}}\widehat{\mathrm{SWD}_1}(X_k,\mathbf{y}) - \max_{k\in\mathcal{A}^c}\widehat{\mathrm{SWD}_1}(X_k,\mathbf{y})\right) > 0$$

(II) Assume Conditions A.2 and B.2 are satisfied. If $d = o\left(n^{q-q\epsilon_0-2}\right)$ for some constant $\epsilon_0 > 0$, then we have almost surely

$$\liminf_{n\to\infty}\left(\min_{k\in\mathcal{A}}\widehat{\mathrm{SWD}_1}(X_k,\mathbf{y}) - \max_{k\in\mathcal{A}^c}\widehat{\mathrm{SWD}_1}(X_k,\mathbf{y})\right) > 0$$

5. Simulations

In this section, we show the empirical performance and the computational time of the proposed SWC screening (SWC-SIS) approach. We compare it with the existing sure independence screening methods, includes sure independence screening (Fan and Lv 2008, SIS), robust rank correlation screening (Li et al. 2012, RRCS), Spearman rank correlation screening (Yan et al. 2017, SRCS), distance correlation based screening (Li et al. 2012, DC-SIS), bias-corrected distance correlation based screening (Székely and Rizzo 2014, bcDC-SIS), martingale difference correlation based screening (Shao and Zhang 2014, MDC-SIS), ball correlation based screening (Pan et al. 2018, BCor-SIS), and projection correlation based method (Liu et al. 2020; Xu et al. 2020, PC-SIS). We consider different data-generating models, including linear models, nonlinear models, and multivariate response models. The model settings are the same as those in Liu et al. (2020). We replicated the experiment a hundred times for each model. For each replicate, we rank the features in descending order w.r.t. each of the feature screening approaches. The screening performance is measured by the following two criteria:

- The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size over 100 replicates.
- The proportion that all active features are selected for a given model size d_i over 100 replicates where $d_1 = [n/\log(n)], d_2 = 2[n/\log(n)], d_3 = 3[n/\log(n)]$.

Throughout this section, we set $n = 100$, $d = 2000$ for each example, and we denote $\Sigma = (\sigma_{ij})_{d\times d}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The numerical results of the second criterion are in Appendix, supplementary materials. What's more, feature screening results for categorically distributed features and response are given in Appendix.

5.1. Linear Models

Let Y be the response variable and X_j be the jth feature of $\mathbf{x}$. Consider the linear model

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + \varepsilon,$$

where the features $\mathbf{x}$ and the random error ε are from the following scenarios.

- **Model 1.a:** $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\varepsilon \sim N(0, 1)$.
- **Model 1.b:** $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\varepsilon \sim t_1(0)$.
- **Model 1.c:** $\mathbf{x} \sim t_1(\mathbf{0}, \mathbf{\Sigma})$ and $\varepsilon \sim N(0, 1)$.
- **Model 1.d:** $\mathbf{x} \sim t_1(\mathbf{0}, \mathbf{\Sigma})$ and $\varepsilon \sim t_1(0)$.

Here, $t_1(0)$ represents for the univariate t-distribution with zero mean and degree-of-freedom one, and $t_1(\mathbf{0}, \mathbf{\Sigma})$ represents multivariate t-distribution with zero mean, degree-of-freedom one, and its variance-covariance matrix is $\mathbf{\Sigma}$. As a result, in Models 1.b–1.d, at least one of the distributions w.r.t. $\mathbf{x}$ and ε is heavy-tailed. We also consider the following two Poisson regression models.

- **Model 1.e** (Continuous): $Y = \exp\{2(X_1 + X_2 + X_3 + X_4 + X_5)\} + \varepsilon$, where $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\varepsilon \sim N(0, 1)$.
- **Model 1.f** (Discrete): $Y \sim \text{Possion}(\exp\{2(X_1 + X_2 + X_3 + X_4 + X_5)\})$, where $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$.

Model 1.e considers the continuous response, while the response in Model 1.f is discrete.

Table 2 summarizes the quantiles of the minimum model size which includes all five active features. We observe that SWC-SIS, PC-SIS, BCor-SIS, SRCS and RRCS perform well under all these linear models, while SIS, MDC-SIS, DC-SIS, and bcDC-SIS suffer from a deteriorated performance at the presence of heavy-tailed features and errors.

5.2. Nonlinear Models

Let Y be the response variable and X_j be the jth feature. Consider the following four nonlinear models

- **Model 2.a:** $Y = 5X_1 + 2\sin(\pi X_2/2) + 2X_3\mathbf{1}_{\{X_3>0\}} + 2\exp\{5X_4\} + \varepsilon$.
- **Model 2.b:** $Y = 3X_1 + 3X_2^3 + 3X_3^{-1} + 5\mathbf{1}_{\{X_4>0\}} + \varepsilon$.
- **Model 2.c:** $Y = 1 - 5(X_2 + X_3)^3 \exp\{-5(X_1 + X_4^3)\} + \varepsilon$.
- **Model 2.d:** $Y = 1 - 5(X_2 + X_3)^{-3} \exp\{1 + 10\sin(\pi X_1/2) + 5X_4\} + \varepsilon$.

where $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\varepsilon \sim N(0, 1)$. Here, we use simple additive structures in Models 2.a and 2.b, and use more challenging nonlinear structures in Models 2.c and 2.d. For each model above, the true model size is 4.

The simulation results are summarized in Table 3. We observe that most of the existing methods suffer from deteriorated

Table 2. The quantiles of minimum model size for linear models over 100 replicates.

	Model 1.a					Model 1.b					Model 1.c				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SWC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.3	54.2	5.0	5.0	5.0	6.0	13.2
PC-SIS	5.0	5.0	5.0	5.0	6.0	5.0	5.0	5.0	7.3	42.1	5.0	5.0	5.0	5.3	12.0
BCor-SIS	5.0	5.0	5.0	5.0	16.3	5.0	5.0	9.0	26.5	112.6	5.0	5.0	7.0	50.0	404.2
MDC-SIS	5.0	5.0	5.0	5.0	5.0	6.0	76.8	569.0	1847.8	1997.1	42.2	327.3	1120.5	1751.0	1972.2
bcDC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.5	11.2	208.5	6.0	32.0	316.0	886.0	1428.6
DC-SIS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	7.0	54.0	463.6	8.0	233.5	965.5	1509.3	1732.1
SRCS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	7.0	33.8	5.0	5.0	5.0	7.0	24.1
RRCS	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	6.6	40.1	5.0	5.0	5.0	6.3	16.2
SIS	5.0	5.0	5.0	5.0	5.0	6.0	89.8	460.5	1508.8	1931.2	113.6	746.3	1514.0	1776.5	1945.4
	Model 1.d					Model 1.e					Model 1.f				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SWC-SIS	5.0	5.0	6.0	11.3	38.1	5.0	5.0	5.0	5.0	6.0	5.0	5.0	5.0	5.0	7.3
PC-SIS	5.0	5.0	6.0	11.0	35.6	5.0	5.0	5.0	5.0	8.0	5.0	5.0	5.0	5.0	7.0
BCor-SIS	5.0	8.8	31.5	121.3	570.9	5.0	5.0	5.0	7.0	18.2	5.0	5.0	5.0	6.0	13.3
MDC-SIS	92.8	623.3	1369.5	1788.5	1967.1	13.4	110.8	329.5	770.0	1198.5	19.9	72.8	164.0	539.5	1244.3
bcDC-SIS	12.9	101.8	461.0	881.8	1374.2	10.8	37.0	108.0	673.5	1778.6	11.0	38.0	84.0	532.3	1602.3
DC-SIS	21.0	329.3	891.0	1431.5	1809.4	32.6	160.5	430.5	711.5	1431.3	41.0	124.3	294.0	558.3	1289.3
SRCS	5.0	5.0	5.0	20.0	69.0	5.0	5.0	5.0	5.0	8.0	5.0	5.0	5.0	5.0	6.0
RRCS	5.0	5.0	7.0	14.0	58.5	5.0	5.0	5.0	5.0	7.1	5.0	5.0	5.0	5.0	6.0
SIS	112.7	951.0	1451.5	1768.8	1921.0	73.9	282.0	539.5	1002.8	1561.5	75.5	181.3	387.5	746.0	1619.1

NOTE: The true model size is 5.

Table 3. The quantiles of minimum model size for nonlinear models over 100 replicates.

	Model 2.a					Model 2.b				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SWC-SIS	4.0	4.0	4.0	5.0	12.1	4.0	4.0	4.0	8.0	102.6
PC-SIS	4.0	4.0	4.0	5.0	9.2	4.0	4.0	4.0	7.0	70.8
BCor-SIS	4.0	4.0	6.0	13.0	80.0	4.0	4.0	6.0	22.5	227.3
MDC-SIS	219.0	727.5	1225.5	1570.0	1593.3	5.0	71.3	985.5	1796.5	1993.1
bcDC-SIS	203.4	944.3	1330.5	1720.8	1945.3	4.0	4.0	5.0	16.0	92.3
DC-SIS	189.6	696.8	1229.5	1588.8	1897.5	4.0	4.8	9.0	78.8	443.2
SRCS	4.0	4.0	5.0	6.0	27.4	4.0	4.0	5.0	9.0	73.2
RRCS	4.0	4.0	4.0	5.0	18.1	4.0	4.0	5.0	9.3	87.2
SIS	218.5	823.8	1273.5	1764.5	1961.3	9.0	117.0	821.0	1569.3	1935.8
	Model 2.c					Model 2.d				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SWC-SIS	4.0	4.0	4.0	5.3	17.1	4.0	4.0	4.0	7.0	21.3
PC-SIS	4.0	4.0	4.0	6.0	16.1	4.0	4.0	7.0	24.0	90.1
BCor-SIS	4.0	4.0	4.0	4.0	6.1	4.0	5.0	7.0	18.0	73.7
MDC-SIS	193.1	1029.0	1535.5	1913.3	2000.0	425.8	1050.8	1574.0	1826.5	1998.2
bcDC-SIS	229.4	1000.5	1217.5	1777.3	1917.7	568.0	1040.8	1447.5	1777.5	1959.3
DC-SIS	323.3	940.0	1373.0	1741.5	1928.7	524.6	1163.3	1532.5	1802.8	1966.6
SRCS	4.0	7.8	24.0	155.4	638.2	4.0	20.0	59.0	341.6	1085.0
RRCS	4.0	7.0	24.5	137.3	634.0	5.0	21.0	69.0	372.8	1205.7
SIS	296.5	934.5	1394.5	1732.3	1921.6	619.2	1237.5	1620.0	1832.3	1950.4

NOTE: The true model size is 4.

Table 4. The quantiles of minimum model size for multivariate response models over 100 replicates.

	Model 3.a					Model 3.b				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SWC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	6.1
PC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	8.3	53.1
BCor-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	6.0	16.5	68.9
MDC-SIS	4.0	4.0	4.0	4.0	4.0	45.9	382.5	814.5	1212.0	1851.7
bcDC-SIS	4.0	4.0	4.0	4.0	4.0	27.6	437.8	1019.5	1412.3	1833.9
DC-SIS	4.0	4.0	4.0	4.0	4.0	172.0	457.0	830.0	1213.3	1782.5
	Model 3.c					Model 3.d				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SWC-SIS	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	5.0
PC-SIS	4.0	4.0	4.0	4.0	4.1	4.0	4.0	5.0	7.0	32.0
BCor-SIS	4.0	4.0	4.0	4.0	30.4	4.0	7.0	18.0	39.3	211.5
MDC-SIS	44.9	275.8	956.5	1436.3	1770.4	770.4	1201.5	1503.5	1795.8	1980.3
bcDC-SIS	10.0	133.3	796.5	1579.0	1883.8	236.2	1201.5	1582.0	1862.0	1978.2
DC-SIS	79.4	269.5	545.0	1016.5	1495.8	356.9	1075.5	1501.5	1701.5	1939.5

NOTE: The true model size is 4.

performance when there exists challenging nonlinear structures. We also observe that the proposed SWC-SIS method performs reasonably well under all nonlinear models.

5.3. Multivariate Response Models

We study the performance of SWC-SIS for multivariate response models. SIS, SRCS, and RRCS are not applicable to multivariate response problems and thus are omitted here. $\mathbf{y} = (Y_1, Y_2)$ are generated from a bivariate normal distribution with conditional mean $\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = (\mu_1(\mathbf{x}), \mu_2(\mathbf{x}))$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} = (\sigma_{ij})_{2\times 2}$, where $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \sigma(\mathbf{x})$. We set $\boldsymbol{\beta} = (\mathbf{2}_4, \mathbf{0}_{d-4})$ and generate $\mu_1(\mathbf{x})$, $\mu_2(\mathbf{x})$ and $\sigma(\mathbf{x})$ from four different models. The first two are borrowed from Li et al. (2012), Liu et al. (2020), and the last two are more complicated.

- **Model 3.a:** $\mu_1(\mathbf{x}) = X_1 + X_3$, $\mu_2(\mathbf{x}) = X_4$ and $\sigma(\mathbf{x}) = \sin(\mathbf{x}\boldsymbol{\beta}^T)$.
- **Model 3.b:** $\mu_1(\mathbf{x}) = X_3\mathbf{1}_{\{X_3>0\}} + \exp\{1 + 10\sin(\pi X_1/2) + 5X_4\}$, $\mu_2(\mathbf{x}) = X_2^{-2}$ and $\sigma(\mathbf{x}) = (\exp\{\mathbf{x}\boldsymbol{\beta}^T\} - 1)/(\exp\{\mathbf{x}\boldsymbol{\beta}^T\} + 1)$.
- **Model 3.c:** $\mu_1(\mathbf{x}) = \exp\{2(X_1 + X_2)\}$, $\mu_2(\mathbf{x}) = X_3 + X_4$ and $\sigma(\mathbf{x}) = \sin(\mathbf{x}\boldsymbol{\beta}^T)$.
- **Model 3.d:** $\mu_1(\mathbf{x}) = 2\sin(\pi X_1/2) + X_3 + \exp\{1 + X_4\}$, $\mu_2(\mathbf{x}) = X_1^{-2} + X_2$ and $\sigma(\mathbf{x}) = (\exp\{\mathbf{x}\boldsymbol{\beta}^T\} - 1)/(\exp\{\mathbf{x}\boldsymbol{\beta}^T\} + 1)$.

In these four models, the union of the active sets of $\mu_1(\mathbf{x})$, $\mu_2(\mathbf{x})$, and $\sigma_1(\mathbf{x})$ contains the first four covariates in the features, and thus the model size is four. The results in Table 4 shows the proposed SWC-SIS method performs reasonably well compared to other methods.

5.4. Computational Time

We compare the CPU time for the proposed SWC method and the competitors. We calculate the dependency between

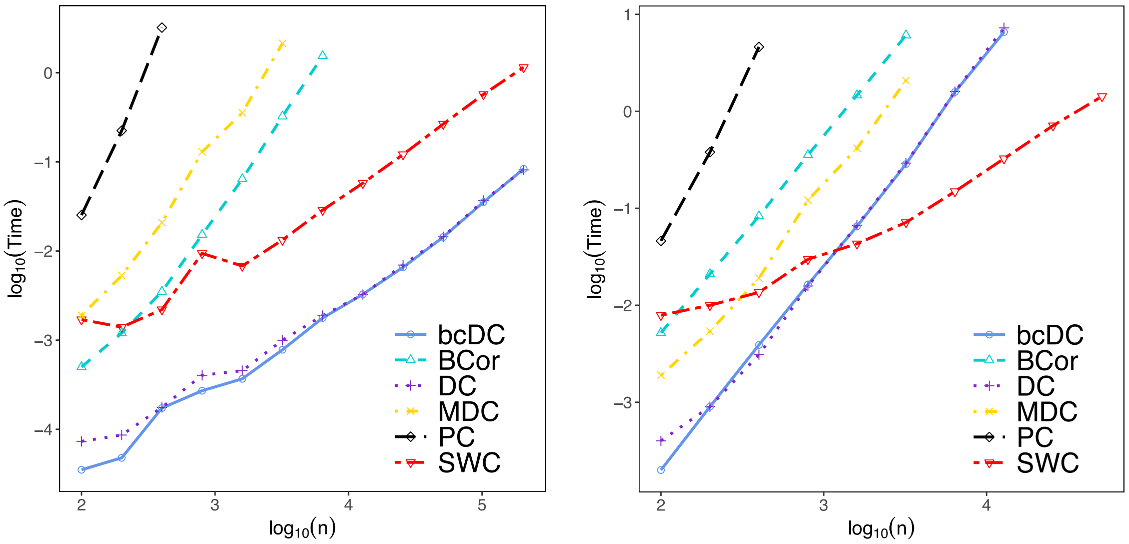

Figure 3. The sample size n versus the CPU time of different dependency measurements for measuring the dependency between two univariate Gaussian variables (left) and two 5d multivariate Gaussian variables.

Table 5. Mean errors of prediction error and time consumption (seconds) based on different feature selection methods for two training sample sizes over 100 replications.

		SWC-SIS	BCor-SIS	MDC-SIS	bcDC-SIS	DC-SIS
30%	Mean error	0.157(0.047)	0.157(0.046)	0.159(0.045)	0.160(0.045)	0.159(0.045)
	Mean screening time(sec)	2.165	2350.210	456.829	9.656	9.651
90%	Mean error	0.048(0.006)	*	*	0.046(0.005)	0.046(0.005)
	Mean screening time(sec)	7.535	*	*	124.640	124.966

two univariate Gaussian random vectors and two multivariate Gaussian random vectors, respectively, with respect to each of the dependency measurements. The average CPU time w.r.t. a hundred replicates are shown in Figure 3, where the left and right panels represent the univariate and multivariate settings, respectively. We first observe that MDC, BCor, and PC method require relatively long CPU time in both settings since the computational cost of these three methods are at least of the order $O(n^2)$. We then observe that DC and bcDC require the shortest CPU time under the univariate setting, while their CPU time significantly increases in the multivariate setting. This observation is expected since the computational cost for these two methods are at the order of $O(n^2)$ in general cases, while they admit efficient algorithm that of the order $O(n\log(n))$ in the univariate setting, see Huo and Székely (2016) for details. Finally, we observe that the CPU time for the proposed SWC method grows approximately linearly w.r.t. n in both settings. Such an observation indicates that SWC is efficient in evaluating the dependency between large-scale random vectors. In addition, the proposed variable screening approach SWC-SIS is efficient for picking information features, especially when the response is multivariate and the sample size n is considerable.

6. Real Data Examples

We consider a multi-response dataset that was first proposed in Spyromitros-Xioufis et al. (2016). This dataset concerns the prediction of river network flows for 48 hr in the future at specific locations. It contains data from hourly flow observations for eight sites in the Mississippi River network in the United States and was obtained from the US National Weather Service. It contains over one year of hourly observations collected from September 2011 to September 2012. Each row includes 576 attribute variables observed from eight sites. Target variables are river network flows for 48 hr in the future of eight sites. The domain is a natural candidate for multi-target regression because there exist clear physical relationships between readings in the contiguous river network (Spyromitros-Xioufis et al. 2016).

After removing missing values, this dataset has sample size $n = 7679$, number of covariates $d = 584$, and number of response $d_y = 8$, with some variables being continuous and some being discrete. We randomly select 30% and 90% of data as the training set, treat the rest as the testing set, and keep $[n_{\text{train}}/\log n_{\text{train}}] = 264$ and $[n_{\text{train}}/(3\log n_{\text{train}})] = 260$ variables respectively in the screening procedures. We use the retained variables to fit a multi-target regression using R package `glmnet`. Finally, we compare mean errors of prediction error based on different feature selection methods over 100 replications. PC-SIS requires huge computational costs, and thus we omit their results here. We summarize the results in Table 5. Both BCor-SIS and MDC-SIS approaches require several hours to calculate the results when using 90% data, and thus their results are omitted. We observe that the proposed SWC-SIS method achieves comparable prediction error with the competitors, requiring significantly less CPU time.

We provide more real-world datasets in supplementary material, including cardiomyopathy microarray data (Segal et al. 2003; Hall and Miller 2009; Li et al. 2012) and yeast cell-cycle data (Chun and Keleş 2010; Chen and Huang 2012; Kong et al. 2017). All the results show that SWC-SIS is effective and efficient for screening informative features for large-scale datasets.

7. Conclusion

We proposed a novel measurement called sliced-Wasserstein dependency to quantify the dependence between two random variables. We then developed a model-free feature screening algorithm by screening out the features whose sliced-Wasserstein dependency w.r.t. the response is relatively small. Theoretically, we showed that our method enjoys sure screening and rank consistency properties under mild regularity conditions. The proposed algorithm is highly efficient for screening informative features in large-scale datasets. The superior performance of our method over mainstream competitors was justified by various numerical experiments.

Supplementary Materials

Appendix: contains the complete proofs of the theoretical results; and additional experiments including two real data examples, simulation results based on the second criterion, and feature screening results for categorically distributed features and response. (appendix.pdf, a pdf file)

Code: contains R code that implements the proposed method and reproduces the numerical results. A readme file is included describing the contents. (code.zip, a zip file)

Acknowledgments

We appreciate the Editor, Associate Editor, and two anonymous reviewers for their constructive comments that helped improve the work.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The authors would like to acknowledge the support from National Natural Science Foundation of China grant No.12101606, No.12001042, No.12271522, and Renmin University of China research fund program for young scholars, and Beijing Institute of Technology research fund program for young scholars.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017), "Wasserstein Generative Adversarial Networks," in *International Conference on Machine Learning*, pp. 214–223. [1503]

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019), "Learning Representations by Maximizing Mutual Information Across Views," *Advances in Neural Information Processing Systems* (Vol. 32). [1503]

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018), "Mutual Information Neural Estimation," in *International Conference on Machine Learning*, pp. 531–540. PMLR. [1503]

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015), "Sliced and Radon Wasserstein Barycenters of Measures," *Journal of Mathematical Imaging and Vision*, 51, 22–45. [1503]

Carriere, M., Cuturi, M., and Oudot, S. (2017), "Sliced Wasserstein Kernel for Persistence Diagrams," in *International Conference on Machine Learning*, pp. 664–673. PMLR. [1502]

Chen, L., and Huang, J. Z. (2012), "Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection," *Journal of the American Statistical Association*, 107, 1533–1545. [1509]

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020), "Faster Wasserstein Distance Estimation with the Sinkhorn Divergence," in *Advances in Neural Information Processing Systems* (Vol. 33), pp. 2257–2269. [1504]

Chun, H., and Keleş, S. (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection," *Journal of the Royal Statistical Society*, Series B, 72, 3–25. [1509]

Cui, H., Li, R., and Zhong, W. (2015), "Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of the American Statistical Association*, 110, 630–641. [1506]

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2022), "False Discovery Rate Control via Data Splitting," *Journal of the American Statistical Association*, 1–38 (just-accepted), DOI: 10.1080/01621459.2022.2060113. [1504]

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. (2019), "Max-Sliced Wasserstein Distance and its Use for Gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656. [1502]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [1501]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1501,1506]

Fan, J., and Ren, Y. (2006), "Statistical Analysis of DNA Microarray Data in Cancer Research," *Clinical Cancer Research*, 12, 4469–4473. [1501]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 2013–2038. [1501]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1501]

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019), "Sample Complexity of Sinkhorn Divergences," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583. PMLR. [1504]

Hall, P., and Miller, H. (2009), "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics*, 18, 533–550. [1501,1509]

He, X., Wang, L., and Hong, H. G. (2013), "Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data," *The Annals of Statistics*, 41, 342–369. [1501]

Huo, X., and Székely, G. J. (2016), "Fast Computing for Distance Covariance," *Technometrics*, 58, 435–447. [1509]

Kantorovich, L. (1942), "On Translation of Mass (in Russian), c r," *Doklady Academy of Sciences of the USSR*, 37, 199–201. [1502]

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019), "Generalized Sliced Wasserstein Distances," in *Advances in Neural Information Processing Systems* (Vol. 32), pp. 261–272. [1502]

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. (2018), "Sliced Wasserstein Auto-Encoders," in *International Conference on Learning Representations*. [1502]

Kolouri, S., Zou, Y., and Rohde, G. K. (2016), "Sliced Wasserstein Kernels for Probability Distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267. [1502]

Kong, Y., Li, D., Fan, Y., and Lv, J. (2017), "Interaction Pursuit in High-Dimensional Multi-Response Regression via Distance Correlation," *The Annals of Statistics*, 45, 897–922. [1509]

Kullback, S. (1997), *Information Theory and Statistics*, Chelmsford, MA: Courier Corporation. [1502]

Levina, E., and Bickel, P. (2001), "The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (Vol. 2), pp. 251–256. IEEE. [1502]

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012), "Robust Rank Correlation based Screening," *The Annals of Statistics*, 40, 1846–1877. [1506]

Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qian, J., Hastie, T., Rivas, M. A., and Tibshirani, R. (2022), "Fast Lasso Method for Large-Scale and

Ultrahigh-Dimensional Cox Model with Applications to UK Biobank," *Biostatistics*, 23, 522–540. [1501]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [1501,1505,1506,1508,1509]

Li, T., Meng, C., Yu, J., and Xu, H. (2022), "Hilbert Curve Projection Distance for Distribution Comparison," arXiv preprint arXiv:2205.15059. [1502]

Liu, J., Li, R., and Wu, R. (2014), "Feature Selection for Varying Coefficient Models with Ultrahigh-Dimensional Covariates," *Journal of the American Statistical Association*, 109, 266–274. [1501]

Liu, J., Zhong, W., and Li, R. (2015), "A Selective Overview of Feature Screening for Ultrahigh-Dimensional Data," *Science China Mathematics*, 58, 1–22. [1501]

Liu, W., Ke, Y., Liu, J., and Li, R. (2020), "Model-Free Feature Screening and FDR Control with Knockoff Features," *Journal of the American Statistical Association*, 117, 428–443. [1501,1505,1506,1508]

Liu, W., and Li, R. (2020), "Variable Selection and Feature Screening," in *Macroeconomic Forecasting in the Era of Big Data* (Vol. 52), ed. P. Fuleky, pp. 293–326, Cham: Springer. [1501]

Lv, J., and Liu, J. S. (2014), "Model Selection Principles in Misspecified Models," *Journal of the Royal Statistical Society*, Series B, 76, 141–167. [1501]

Mai, Q., and Zou, H. (2015), "The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method," *The Annals of Statistics*, 43, 1471–1497. [1501]

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., and Ma, P. (2019), "Large-Scale Optimal Transport Map Estimation Using Projection Pursuit," in *Advances in Neural Information Processing Systems* (Vol. 32). [1502]

Mordant, G., and Segers, J. (2022), "Measuring Dependence between Random Vectors via Optimal Transport," *Journal of Multivariate Analysis*, 189, 104912. [1503]

Nadjahi, K. (2021), "Sliced-Wasserstein Distance for Large-Scale Machine Learning: Theory, Methodology and Extensions," Ph. D. thesis, Institut polytechnique de Paris. [1502]

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., and Şimşekli, U. (2020), "Approximate Bayesian Computation with the Sliced-Wasserstein Distance," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5470–5474. IEEE. [1502]

Nguyen, K., Ho, N., Pham, T., and Bui, H. (2020), "Distributional Sliced-Wasserstein and Applications to Generative Modeling," arXiv preprint arXiv:2002.07367. [1502]

Nies, T. G., Staudt, T., and Munk, A. (2021), "Transport Dependency: Optimal Transport based Dependency Measures," arXiv preprint arXiv:2105.02073. [1503]

Ozair, S., Lynch, C., Bengio, Y., van den Oord, A., Levine, S., and Sermanet, P. (2019), "Wasserstein Dependency Measure for Representation Learning," *Advances in Neural Information Processing Systems* (Vol. 32), pp. 15604–15614. [1503]

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018), "A Generic Sure Independence Screening Procedure," *Journal of the American Statistical Association*, 928–937. [1506]

Panaretos, V. M., and Zemel, Y. (2019), "Statistical Aspects of Wasserstein Distances," *Annual Review of Statistics and its Application*, 6, 405–431. [1502]

Pang, G., Cao, L., Chen, L., and Liu, H. (2018), "Learning Representations of Ultrahigh-Dimensional Data for Random Distance-based Outlier Detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2041–2050. [1501]

Peng, H., Long, F., and Ding, C. (2005), "Feature Selection based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1226–1238. [1503]

Peyré, G., and Cuturi, M. (2019), "Computational Optimal Transport," *Foundations and Trends® in Machine Learning*, 11, 355–607. [1502,1504]

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011), "Wasserstein Barycenter and its Application to Texture Mixing," in *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446, Springer. [1502,1503]

Rowland, M., Hron, J., Tang, Y., Choromanski, K., Sarlós, T., and Weller, A. (2019), "Orthogonal Estimation of Wasserstein Distances," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 186–195. PMLR. [1502]

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression Approaches for Microarray Data Analysis," *Journal of Computational Biology*, 10, 961–980. [1509]

Shao, X., and Zhang, J. (2014), "Martingale Difference Correlation and its Use in High-Dimensional Variable Screening," *Journal of the American Statistical Association*, 109, 1302–1318. [1506]

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., and Vlahavas, I. (2016), "Multi-Target Regression via Input Space Expansion: Treating Targets as Inputs," *Machine Learning*, 104, 55–98. [1509]

Stone, J. V. (2004), *Independent Component Analysis: A Tutorial Introduction*, Cambridge, MA: MIT Press. [1503]

Székely, G. J., and Rizzo, M. L. (2014), "Partial Distance Correlation with Methods for Dissimilarities," *The Annals of Statistics*, 42, 2382–2412. [1506]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [1503]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003), "Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays," *Statistical Science*, 104–117. [1501]

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009), "Feature Hashing for Large Scale Multitask Learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1113–1120. [1501]

Wiesel, J. C. (2022), "Measuring Association with Wasserstein Distances," *Bernoulli*, 28, 2816–2832. [1503]

Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., and Gool, L. V. (2019), "Sliced Wasserstein Generative Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722. [1502]

Wu, Y., and Yin, G. (2015), "Conditional Quantile Screening in Ultrahigh-Dimensional Heterogeneous Data," *Biometrika*, 102, 65–76. [1501]

Xu, H., Luo, D., Henao, R., Shah, S., and Carin, L. (2020), "Learning Autoencoders with Relational Regularization," in *International Conference on Machine Learning*, pp. 10576–10586. PMLR. [1502]

Xu, K., Shen, Z., Huang, X., and Cheng, Q. (2020), "Projection Correlation between Scalar and Vector Variables and its Use in Feature Screening with Multi-Response Data," *Journal of Statistical Computation and Simulation*, 90, 1923–1942. [1501,1506]

Xue, L., and Zou, H. (2011), "Sure Independence Screening and Compressed Random Sensing," *Biometrika*, 98, 371–380. [1501]

Yan, X., Tang, N., and Zhao, X. (2017), "The Spearman Rank Correlation Screening for Ultrahigh Dimensional Censored Data," arXiv preprint arXiv:1702.02708. [1506]

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011), "Model-Free Feature Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1475. [1501]