KPOTD: Kernel Principal Optimal Transport Directions for Nonlinear Sufficient Dimension Reduction

Peize Wang* **Wenhao Jiang*** **Cheng Meng†**
Institute of Statistics and Big Data, Renmin University of China
Beijing, China
wpz2024@ruc.edu.cn; 2021201382@ruc.edu.cn; chengmeng@ruc.edu.cn
*Equal contribution. †Corresponding author.

Abstract

Nonlinear sufficient dimension reduction (SDR) is a fundamental tool for identifying complex low-dimensional structures that preserve the conditional distribution of a response variable. Existing nonlinear SDR paradigms primarily rely on inverse conditional moments, large margins, or deep learning. However, these approaches often struggle with highly symmetric distributions, focus on localized boundaries, or lack statistical interpretability. In this paper, we propose a new framework named Kernel Principal Optimal Transport Directions (KPOTD). By integrating optimal transport into the RKHS, KPOTD constructs a displacement covariance operator that aligns probability measures globally. This approach allows the method to robustly untangle complex nonlinear manifolds and achieve linear separability in the latent space. Computationally, KPOTD reduces the infinite-dimensional functional optimization to a finite-dimensional generalized eigenvalue problem. Theoretically, we investigate the properties of the KPOTD operator and rigorously establish its unbiasedness and exhaustiveness. These results guarantee the exact recovery of the central subspace without relying on restrictive classical assumptions. Extensive numerical simulations and real-data applications demonstrate that KPOTD consistently outperforms existing methods and maintains high stability even in the presence of high-dimensional noise.

1 Introduction

Sufficient dimension reduction (SDR) addresses the curse of dimensionality by identifying a central subspace that preserves the conditional distribution of a response variable without parametric assumptions. However, classical SDR is fundamentally restricted to linear projections, rendering it inadequate for capturing the complex, nonlinear manifolds inherent in modern high-dimensional data. To overcome this limitation, extensions to nonlinear functional spaces typically follow three main paradigms: inverse conditional moments [Lee et al., 2013, Wu, 2008], large-margin classification [Li et al., 2011], and, more recently, deep-learning-based architectures [Liang et al., 2022, Huang et al., 2024]. However, moment-based methods primarily capture lower-order statistics and often fail on highly symmetric distributions, while margin-based approaches focus on localized boundaries rather than overall topological structures. Similarly, while deep-learning methods offer immense representational power, they often lack statistical interpretability and require extensive hyperparameter tuning [Chen et al., 2025, Liang et al., 2022]. While optimal transport (OT) provides a global structural alignment mechanism that overcomes localized boundary issues, as demonstrated by the linear Principal Optimal Transport Direction (POTD) [Meng et al., 2020], its linear formulation restricts its applicability on complex nonlinear manifolds. This motivates the need for a framework

STAI-X 2026 (Full Paper Track).

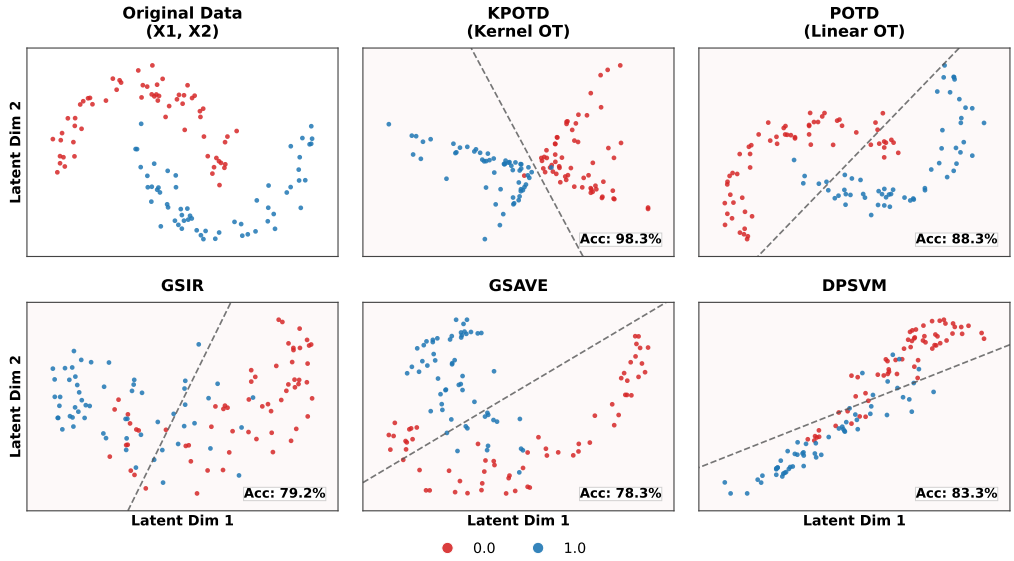

Figure 1: **Latent Representation Recovery vs. Linear Subspace Projection on Entangled Nonlinear Topologies. Setting:** The dataset consists of samples from two entangled manifolds embedded in a 10-dimensional space with 8 nuisance variables ($p = 10, d = 2$). **Top Row:** Given the original data (left), **POTD** (right) identifies the correct subspace but fails to achieve linear separability. Our proposed **KPOTD** (center) transforms the nonlinear manifold into a separable latent space (98.3% linear classification accuracy). **Bottom Row:** Comparisons with moment-based (**GSIR**, **GSAVE**) and margin-based (**DPSVM**) methods, which fail to recover directions on symmetric distributions or lack overall topological coherence.

that combines the global geometric alignment of OT with the flexible representational capacity of functional spaces.

To formally illustrate this limitation, consider a motivating example where the active coordinates $X_{\mathcal{A}} \in \mathbb{R}^2$ follow an entangled nonlinear topology: $X_{\mathcal{A},y=0} = (\cos\theta, \sin\theta)^\top + \varepsilon$ and $X_{\mathcal{A},y=1} = (1 - \cos\theta, 0.5 - \sin\theta)^\top + \varepsilon$ for $\theta \in [0, \pi]$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_2)$. We embed $X_{\mathcal{A}}$ into a high-dimensional space ($p = 10$) with 8 Gaussian nuisance variables. As illustrated in Figure 1, linear POTD merely identifies a Euclidean subspace; the nonlinear entanglement persists, leaving the latent representations linearly inseparable.

To address these limitations, we propose Kernel Principal Optimal Transport Directions (KPOTD). By constructing an OT displacement covariance operator within an RKHS, KPOTD maps nonlinear manifolds into linearly separable representations (Figure 1). This global alignment mechanism bypasses existing constraints and remains robust under high-dimensional noise.

Our contributions are summarized as follows:

- We develop KPOTD, a novel nonlinear dimension reduction framework that overcomes the theoretical and computational intractability of optimal transport in infinite-dimensional functional spaces. Our core methodological innovation is the formulation of a second-order OT displacement covariance operator within an RKHS, which captures global structural discrepancies to transform entangled manifolds into separable representations.
- We devise a stable numerical implementation that reduces the infinite-dimensional functional optimization to a finite-dimensional generalized eigenvalue problem. This algebraic design inherently obviates the need for explicit boundary calculations, ensuring a highly robust and computationally efficient solution for high-dimensional data.
- We establish the theoretical foundations of the KPOTD operator by rigorously proving its unbiasedness and exhaustiveness, ensuring the exact recovery of the nonlinear central subspace at the population level under very mild conditions.

- We conduct extensive experiments on complex manifolds and real-world benchmarks to demonstrate the superior learning capability of KPOTD. Our method exhibits exceptional robustness to high-dimensional noise, consistently outperforming both traditional linear and modern nonlinear baselines in classification accuracy across diverse datasets.

2 Preliminaries

In this section, we review the fundamental concepts that form the mathematical foundation of our proposed method. We introduce the formulation of sufficient dimension reduction (SDR), its nonlinear extension to central subspaces, and finally, the mathematical formulation of optimal transport (OT) and POTD within functional spaces.

2.1 Sufficient Dimension Reduction and Nonlinear Central Subspace

Let $\boldsymbol{X} \in \mathbb{R}^p$ be a high-dimensional predictor vector and $Y \in \{1, \ldots, k\}$ be the corresponding categorical response variable. Linear SDR seeks a projection matrix $\boldsymbol{B} \in \mathbb{R}^{p\times r}$ with $r \ll p$, such that the conditional distribution of Y given $\boldsymbol{X}$ is completely preserved by the projected predictors $\boldsymbol{B}^\top \boldsymbol{X}$. This is formalized by the conditional independence condition:

$$Y \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{B}^\top \boldsymbol{X}. \tag{1}$$

The intersection of all such dimension reduction subspaces is defined as the linear central subspace, denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$ [Cook, 1998].

To capture nonlinear dependencies, nonlinear SDR can be formulated through σ-fields [Lee et al., 2013]: an SDR σ-field for Y versus X is a sub-σ-field $\mathcal{G} \subseteq \sigma(X)$ satisfying $Y \perp\!\!\!\perp X \mid \mathcal{G}$. When it exists, the minimal such σ-field is the central σ-field, and its square-integrable measurable functions form the central class. Since KPOTD constructs directions in the RKHS through the feature map $\phi : \mathcal{X} \to \mathcal{H}$, we represent the sufficient information by the projection-generated σ-field $\sigma\{P_{\mathcal{S}}\phi(\boldsymbol{X})\}$, where $\mathcal{S} \subseteq \mathcal{H}$ is a finite-dimensional closed subspace, and seek $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp \phi(\boldsymbol{X}) \mid P_{\mathcal{S}}\phi(\boldsymbol{X}). \tag{2}$$

When a minimal such subspace exists, we denote it by $\mathcal{S}_{Y|\phi(\boldsymbol{X})}$. A more detailed discussion is given in Appendix A.5.

2.2 Optimal Transport and Principal Optimal Transport Direction

Optimal transport (OT) provides a mathematical framework to define distances between probability measures [Villani, 2009]. OT has recently emerged as a powerful framework across various statistical learning domains, supported by scalable map estimation techniques [Meng et al., 2019] and comprehensive overviews of modern computational OT methods [Zhang et al., 2021, 2023]. Beyond dimension reduction, OT-based discrepancy measures have been successfully applied to related tasks such as model-free feature screening [Li et al., 2023c].

For two probability measures μ and ν on a space Ω, the Kantorovich formulation seeks a joint coupling π that minimizes the expected transportation cost $c(\boldsymbol{x}, \boldsymbol{y})$:

$$W_c(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} c(\boldsymbol{x}, \boldsymbol{y}) d\pi(\boldsymbol{x}, \boldsymbol{y}), \tag{3}$$

where $\Pi(\mu, \nu)$ denotes the set of all joint distributions with marginals μ and ν [Peyré and Cuturi, 2019].

2.3 Optimal Transport in RKHS

Let $\phi_{\#}\mu$ and $\phi_{\#}\nu$ denote the pushforward probability measures of μ and ν induced by the feature map $\phi : \mathcal{X} \to \mathcal{H}$. The Kantorovich optimal transport problem between these pushforward measures in the RKHS is defined as

$$W^2_{\mathcal{H}}(\phi_{\#}\mu, \phi_{\#}\nu) = \inf_{\pi_{\mathcal{H}}\in\Pi(\phi_{\#}\mu,\phi_{\#}\nu)} \int_{\mathcal{H}\times\mathcal{H}} \|u - v\|^2_{\mathcal{H}} d\pi_{\mathcal{H}}(u, v), \tag{4}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm induced by the inner product in $\mathcal{H}$.

To alleviate the computational burden associated with exact Wasserstein distances, recent studies have developed efficient surrogates, such as Hilbert curve projection distances [Li et al., 2024] and space-filling curve projections [Li et al., 2025]. Although $\mathcal{H}$ can be infinite-dimensional, this functional optimal transport problem can be exactly evaluated via the kernel trick [Zhang et al., 2020]. Because the optimal coupling $\pi^*_{\mathcal{H}}$ and the corresponding functional displacements depend strictly on data inner products, they are directly computable using the kernel $k(\boldsymbol{x}, \boldsymbol{y})$. This establishes the mathematical foundation for the proposed KPOTD method detailed in Section 3.

3 Methodology

In this section, we develop the proposed framework. We first formulate the optimal transport displacement covariance operator in an RKHS. Subsequently, we detail the empirical estimation of the operator and demonstrate the reduction of the functional optimization to a finite-dimensional generalized eigenvalue problem.

The Population-Level Operator. Let $\phi : \mathcal{X} \to \mathcal{H}$ be a feature map to an RKHS equipped with a positive definite kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_{\mathcal{H}}$. For a categorical response $Y \in \{1, \ldots, k\}$ with prior probabilities $p_i = \mathbb{P}(Y = i)$, let $\boldsymbol{\pi}^*_{ij} \in \Pi(\mathcal{L}(X \mid Y = i), \mathcal{L}(X \mid Y = j))$ denote the optimal transport plan between the conditional distributions of $X \mid Y = i$ and $X \mid Y = j$ under the RKHS-induced quadratic cost $c_{\mathcal{H}}(x, x') = \|\phi(x) - \phi(x')\|^2_{\mathcal{H}}$. We define the population-level OT displacement covariance operator $\Sigma_{\mathrm{OT}} : \mathcal{H} \to \mathcal{H}$ as:

$$\Sigma_{\mathrm{OT}} = \sum_{1 \le i < j \le k} \omega_{ij} \mathbb{E}_{\boldsymbol{\pi}^*_{ij}} \left[(\phi(\boldsymbol{X}_i) - \phi(\boldsymbol{X}_j)) \otimes (\phi(\boldsymbol{X}_i) - \phi(\boldsymbol{X}_j)) \right], \tag{5}$$

where $\otimes$ denotes the tensor product, and ω_{ij} is the pairwise weight. While setting $\omega_{ij} = p_i p_j$ is a natural choice based on prior probabilities, in practice, we fix $\omega_{ij} = 1$ to apply uniform weighting. KPOTD seeks an orthogonal projection operator onto an r-dimensional subspace in $\mathcal{H}$ that maximizes the projected displacement variance, formulated as:

$$\max_{f_1, \ldots, f_r} \sum_{m=1}^{r} \langle f_m, \Sigma_{\mathrm{OT}} f_m \rangle_{\mathcal{H}} \quad \text{s.t.} \quad \langle f_m, f_l \rangle_{\mathcal{H}} = \delta_{ml}. \tag{6}$$

Empirical Estimation. To estimate the operator empirically given sample matrices $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, we first compute the pairwise squared distance matrix $\boldsymbol{C}$ in $\mathcal{H}$ directly via the kernel trick, where $\boldsymbol{C}_{u,v} = k(\boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,u}) + k(\boldsymbol{x}_{j,v}, \boldsymbol{x}_{j,v}) - 2k(\boldsymbol{x}_{i,u}, \boldsymbol{x}_{j,v})$. Using uniform marginal weights, the discrete optimal transport plan $\boldsymbol{\pi}^*_{ij}$ is obtained by solving the standard Kantorovich transportation problem over $\boldsymbol{C}$.

By the Representer Theorem, the optimal eigenfunctions in Equation 6 lie within the span of the empirically mapped data: $f(\cdot) = \sum_{m=1}^{n} \boldsymbol{\alpha}_m \phi(\boldsymbol{x}_m)$, where $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $n = \sum_{i=1}^{k} n_i$. Substituting this expansion into the empirical objective transforms the functional optimization into a finite-dimensional generalized eigenvalue problem:

$$\boldsymbol{K}\boldsymbol{M}\boldsymbol{K}\boldsymbol{\alpha} = \lambda(\boldsymbol{K} + \gamma \boldsymbol{I})\boldsymbol{\alpha}, \tag{7}$$

where $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is the kernel Gram matrix evaluated on sample, $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is a symmetric block matrix constructed from the optimal transport couplings $\{\boldsymbol{\pi}^*_{ij}\}$, and $\gamma > 0$ is a Tikhonov regularization parameter to ensure numerical stability. The derivation of $\boldsymbol{M}$ is provided in Appendix C, and the overall procedure is summarized in Algorithm 1.

Dimension Selection and Complexity. The structural dimension r is determined by the spectral properties of the generalized eigenvalue problem. Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$ be the sorted eigenvalues. The dimension r is chosen via the cumulative eigenvalue ratio:

$$r = \min \left\{ d : \frac{\sum_{m=1}^{d} \lambda_m}{\sum_{m=1}^{n} \lambda_m} \ge \eta \right\}, \tag{8}$$

where $\eta \in (0, 1)$ is a predefined threshold (e.g., 0.90 or 0.95).

The computational complexity of KPOTD involves three main steps: computing the kernel matrix ($\mathcal{O}(n^2p)$), solving the regularized optimal transport via Sinkhorn iterations ($\mathcal{O}(Ln^2)$, with L iterations), and solving the generalized eigenvalue problem ($\mathcal{O}(n^3)$). The overall time complexity is $\mathcal{O}(n^3 + n^2p)$. While standard Sinkhorn iterations require $\mathcal{O}(Ln^2)$ time, the computational bottleneck of large-scale optimal transport can be significantly alleviated using recent sparsification techniques. For instance, importance sparsification methods and related sampling-based approaches can effectively reduce the complexity of entropic OT to near-linear time by preserving only the most informative entries in the cost matrix [Li et al., 2023a,b, Hu et al., 2025, Ouyang et al., 2026]. For large-scale applications, approximation techniques such as the Nyström method or Random Fourier Features can be employed to reduce the $\mathcal{O}(n^3)$ cost to $\mathcal{O}(nd^2)$, where $d \ll n$.

Algorithm 1 Multiclass Kernel Principal Optimal Transport Direction (KPOTD)

Input: Data grouped by class $\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(k)}$, target dimension r, kernel matrix $\mathbf{K}$, ridge γ, marginals $\mathbf{a}_1, \ldots, \mathbf{a}_k$.

for $i = 1, \ldots, k$ **do**
  for $j = i+1, \ldots, k$ **do**
    $\mathbf{C} \leftarrow \text{diag}(\mathbf{K}_{ii})\mathbf{1}^\top + \mathbf{1}\text{diag}(\mathbf{K}_{jj})^\top - 2\mathbf{K}_{ij}$
    $\boldsymbol{\pi}_{ij} \leftarrow \text{OT}[(\mathbf{X}_{(i)}, \mathbf{a}_i), (\mathbf{X}_{(j)}, \mathbf{a}_j), \text{cost} = \mathbf{C}]$
  end for
end for

Assemble the global block matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$:

$$\mathbf{M} = \begin{pmatrix} (k-1)\text{diag}(\mathbf{a}_1) & -\boldsymbol{\pi}_{12} & \cdots & -\boldsymbol{\pi}_{1k} \\ -\boldsymbol{\pi}_{12}^\top & (k-1)\text{diag}(\mathbf{a}_2) & \cdots & -\boldsymbol{\pi}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{\pi}_{1k}^\top & -\boldsymbol{\pi}_{2k}^\top & \cdots & (k-1)\text{diag}(\mathbf{a}_k) \end{pmatrix}$$

Solve the generalized eigenvalue problem:

$$\mathbf{KMK}\boldsymbol{\alpha} = \lambda(\mathbf{K} + \gamma\mathbf{I})\boldsymbol{\alpha}$$

Let $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r]$ be the top r eigenvectors.
Normalize each $\boldsymbol{\alpha}_m$ such that $\boldsymbol{\alpha}_m^\top \mathbf{K} \boldsymbol{\alpha}_m = 1$.

Output: Coefficient matrix $\mathbf{A}$

4 Theoretical Properties

Let $\mathcal{H}$ be the RKHS induced by a kernel k, and let $\phi : \mathcal{X} \to \mathcal{H}$ be the corresponding feature map. For a closed subspace $\mathcal{S} \subseteq \mathcal{H}$, denote by $P_{\mathcal{S}}$ and $P_{\mathcal{S}^\perp}$ the orthogonal projections onto $\mathcal{S}$ and its orthogonal complement.

Assumption 1 (Finite-dimensional nonlinear central subspace)**.** There exists a finite-dimensional closed subspace $\mathcal{S}_{Y|\phi(\boldsymbol{X})} \subseteq \mathcal{H}$ such that $Y \perp\!\!\!\perp \phi(\boldsymbol{X}) \mid P_{\mathcal{S}_{Y|\phi(\boldsymbol{X})}}\phi(\boldsymbol{X})$, and $\mathcal{S}_{Y|\phi(\boldsymbol{X})}$ is minimal among closed subspaces satisfying this conditional independence relation. Throughout this section, we write $\mathcal{S} = \mathcal{S}_{Y|\phi(\boldsymbol{X})}$.

Assumption 2 (Pairwise OT reducibility)**.** Let $\boldsymbol{Z} = P_{\mathcal{S}}\phi(\boldsymbol{X})$ and $\boldsymbol{N} = P_{\mathcal{S}^\perp}\phi(\boldsymbol{X})$. For each class pair (i, j), define $\mu_i = \mathcal{L}(\phi(\boldsymbol{X}) \mid Y = i)$, $\mu_j = \mathcal{L}(\phi(\boldsymbol{X}) \mid Y = j)$, and $\nu_i = \mathcal{L}(\boldsymbol{Z} \mid Y = i)$, $\nu_j = \mathcal{L}(\boldsymbol{Z} \mid Y = j)$. We assume that the quadratic Wasserstein discrepancy between the full class-conditional RKHS distributions is completely determined by their signal projections: $W_2^2(\mu_i, \mu_j) = W_2^2(\nu_i, \nu_j)$, where W_2 on $\mathcal{H}$ is computed using the Hilbert norm $\|\cdot\|_{\mathcal{H}}$, and W_2 on $\mathcal{S}$ is computed using the restricted Hilbert norm.

Remark 1. Assumption 2 means that the optimal-transport discrepancy between two class-conditional RKHS distributions is attributable to the signal component rather than the nuisance variation. It ensures the class-pair OT displacements can be interpreted as genuine class-to-class signal movements. A sufficient signal-plus-nuisance generative example satisfying this reducibility condition, and also the separation condition used for exhaustiveness below, is provided in Appendix A.6.

Lemma 1 (Basic properties of the OT displacement operator)**.** *Suppose that $\mu_i, \mu_j \in \mathcal{P}_2(\mathcal{H})$ for every class pair (i, j), and that the class-pair weights satisfy $\omega_{ij} \geq 0$. Then Σ_{OT} is a bounded, self-adjoint, positive semidefinite, trace-class operator on $\mathcal{H}$. In particular, Σ_{OT} is compact.*

Lemma 2 (No nuisance displacement under OT reducibility)**.** *Suppose Assumption 2 holds. Then, for any class pair (i, j), every quadratic-cost optimal coupling $\boldsymbol{\pi}_{ij}^* \in \Pi(\mu_i, \mu_j)$ satisfies*

$$P_{\mathcal{S}^\perp}\xi = P_{\mathcal{S}^\perp}\zeta \quad \boldsymbol{\pi}_{ij}^*\text{-a.s.}$$

Theorem 3 (Unbiasedness of the population KPOTD operator)**.** *Suppose Assumptions 1 and 2 hold. Then*

$$P_{\mathcal{S}^\perp}\Sigma_{\mathrm{OT}} = 0, \qquad \overline{\mathrm{ran}(\Sigma_{\mathrm{OT}})} \subseteq \mathcal{S}.$$

Consequently, every population KPOTD direction extracted from the leading eigenspaces of Σ_{OT} belongs to the nonlinear central subspace $\mathcal{S}$. Thus, the population KPOTD operator is unbiased.

The proofs of Lemmas 1 and 2, together with the proof of Theorem 3, are provided in Appendix A.3.

Building upon unbiasedness, we next give a sufficient condition under which the population OT displacement operator is exhaustive. Intuitively, unbiasedness rules out spurious directions in $\mathcal{S}^\perp$, while exhaustiveness requires that no nonzero direction inside $\mathcal{S}$ is missed by the class-conditional OT displacements.

Assumption 3 (Nondegenerate OT separation)**.** For the optimal couplings $\{\boldsymbol{\pi}_{ij}^*\}$ used in the definition of Σ_{OT}, every nonzero direction $u \in \mathcal{S}$ satisfies $\sum_{i<j} \omega_{ij} \int \langle u, \xi - \zeta \rangle_{\mathcal{H}}^2 \, d\boldsymbol{\pi}_{ij}^*(\xi, \zeta) > 0$.

Assumption 3 states that every direction in the nonlinear central subspace is detected by at least one weighted class-pair OT displacement. Equivalently, the restriction of Σ_{OT} to $\mathcal{S}$ has no nontrivial null direction.

Theorem 4 (Exhaustiveness of the population KPOTD operator)**.** *Suppose the conditions in Theorem 3 hold. If, in addition, Assumption 3 holds, then*

$$\overline{\mathrm{ran}(\Sigma_{\mathrm{OT}})} = \mathcal{S}.$$

Thus, the population KPOTD operator exhaustively recovers the nonlinear central subspace.

The proof of Theorem 4 is provided in Appendix A.4.

Implication. Theorems 3 and 4 show that, at the population level, the KPOTD displacement operator identifies the finite-dimensional RKHS subspace $\mathcal{S}$ whose projection $P_{\mathcal{S}}\phi(\boldsymbol{X})$ is sufficient for Y. Under OT reducibility, the class-pair displacement operator has no energy in $\mathcal{S}^\perp$; under nondegenerate OT separation, it has no missing direction inside $\mathcal{S}$. Thus KPOTD provides a global distributional mechanism for recovering the projection-generated central subspace, complementing moment-based and margin-based nonlinear SDR methods.

Convergence analysis. We further analyze the finite-sample behavior of the empirical KPOTD operator through an operator-level perturbation framework and discuss its implications for eigenspace convergence under fixed or empirical class-pair weights; the technical statements and proofs, incorporating the empirical coupling $\widehat{\boldsymbol{\pi}}_{ij}$ and the population coupling $\boldsymbol{\pi}_{ij}^*$, are deferred to Appendix B.

5 Numerical Experiments

We evaluate KPOTD against a comprehensive suite of linear and nonlinear SDR baselines. Section 5.1 uses synthetic manifolds to test each method's ability to capture nonlinear geometric structures, while Section 5.2 assesses practical scalability across real-world classification tasks. In all experiments, representation quality and computational efficiency are measured by downstream k-nearest-neighbor accuracy and running time, respectively.

5.1 Simulation Studies

We evaluate the finite-sample performance of KPOTD against several representative nonlinear SDR baselines, including generalized sliced inverse-regression and average-variance estimators (GSIR

Table 1: Average downstream classification accuracies (standard errors) of different methods across four DGPs with varying feature dimensions $p \in \{10, 20, 30\}$ over 100 simulation runs. The downstream classifier is 5-NN. Bold-faced numbers indicate the best performers.

p	Method	Model I	Model II	Model III	Model IV
10	**KPOTD (Ours)**	**0.9399 (0.0029)**	**0.9410 (0.0022)**	**0.9338 (0.0025)**	**0.9503 (0.0020)**
	GSIR	0.8863 (0.0054)	0.9163 (0.0029)	0.8562 (0.0063)	0.7519 (0.0101)
	GSAVE	0.9265 (0.0030)	0.8749 (0.0045)	0.9126 (0.0039)	0.9421 (0.0018)
	KSIR	0.5224(0.0043)	0.8052 (0.0034)	0.5607(0.0064)	0.5336 (0.0072)
	KPSVM	0.7765 (0.0060)	0.7709 (0.0046)	0.8078 (0.0052)	0.8397 (0.0069)
	dCov-SDR	0.9143 (0.0070)	0.8020 (0.0039)	0.9255 (0.0046)	0.9478 (0.0021)
	KDR	0.6003 (0.0015)	0.8522 (0.0038)	0.5600 (0.0154)	0.5313 (0.0256)
	NN	0.8754 (0.0038)	0.8353 (0.0051)	0.8782 (0.0044)	0.9161 (0.0035)
	StoNet	0.5145(0.0043)	0.7349 (0.0038)	0.6482 (0.0044)	0.5567 (0.0046)
	DPSVM	0.8011(0.0039)	0.8772 (0.0032)	0.8763 (0.0039)	0.9074(0.0028)
	DDR	0.7930 (0.0071)	0.8180 (0.0037)	0.8145 (0.0054)	0.8847 (0.0047)
20	**KPOTD (Ours)**	**0.9342 (0.0027)**	**0.9407 (0.0022)**	**0.9232 (0.0026)**	**0.9490 (0.0022)**
	GSIR	0.8743 (0.0052)	0.9128 (0.0034)	0.8067 (0.0068)	0.6548 (0.0101)
	GSAVE	0.9238 (0.0027)	0.8818 (0.0043)	0.8950 (0.0047)	0.9192 (0.0033)
	KSIR	0.5082 (0.0044)	0.7751 (0.0040)	0.5232 (0.0047)	0.5042 (0.0048)
	KPSVM	0.7365 (0.0048)	0.7600 (0.0046)	0.7304 (0.0044)	0.6269 (0.0059)
	dCov-SDR	0.8000 (0.0136)	0.7715 (0.0041)	0.7814 (0.0115)	0.9296 (0.0031)
	KDR	0.5000 (0.0055)	0.8045 (0.0044)	0.5089 (0.0065)	0.4891 (0.0201)
	NN	0.8340 (0.0038)	0.7992 (0.0052)	0.8057 (0.0045)	0.8771 (0.0050)
	StoNet	0.5052(0.0045)	0.6847 (0.0049)	0.4993 (0.0035)	0.4971 (0.0043)
	DPSVM	0.7674(0.0051)	0.8758 (0.0038)	0.7838 (0.0046)	0.8760 (0.0031)
	DDR	0.6648 (0.0063)	0.8557 (0.0035)	0.6996 (0.0060)	0.6807 (0.0109)
30	**KPOTD (Ours)**	**0.9393 (0.0022)**	**0.9402 (0.0023)**	**0.9174 (0.0027)**	**0.9436 (0.0021)**
	GSIR	0.8575(0.0061)	0.9142 (0.0032)	0.7698 (0.0073)	0.5892 (0.0104)
	GSAVE	0.9227 (0.0031)	0.8635 (0.0037)	0.8930 (0.0034)	0.8763 (0.0054)
	KSIR	0.5034 (0.0045)	0.7490 (0.0059)	0.5020 (0.0049)	0.5021 (0.0044)
	KPSVM	0.7165 (0.0046)	0.7503 (0.0046)	0.7141 (0.0038)	0.5451 (0.0053)
	dCov-SDR	0.6755 (0.0131)	0.7578 (0.0039)	0.6407 (0.0086)	0.8684 (0.0078)
	KDR	0.4950 (0.0044)	0.7212 (0.0048)	0.4960 (0.0046)	0.5026 (0.0048)
	NN	0.7770 (0.0052)	0.7850 (0.0051)	0.7437 (0.0043)	0.8086 (0.0063)
	StoNet	0.5051(0.0051)	0.6722 (0.0042)	0.5062 (0.0045)	0.5058 (0.0046)
	DPSVM	0.7197 (0.0046)	0.8941 (0.0028)	0.7343 (0.0043)	0.7502 (0.0048)
	DDR	0.6082 (0.0056)	0.7685 (0.0036)	0.6407 (0.0058)	0.5716 (0.0071)

and GSAVE [Lee et al., 2013]), kernel/RKHS-based methods (KSIR [Wu, 2008], KDR [Fukumizu et al., 2004]), margin-based kernel SDR (KPSVM [Li et al., 2011]), dependence-based SDR (dCov-SDR [Sheng and Yin, 2013]), and neural-network-based methods (NN [Kapla et al., 2022], DPSVM [Chen et al., 2025], DDR [Huang et al., 2024], StoNet [Liang et al., 2022]). Hyperparameters are selected by 3-fold cross-validation when method-specific defaults are unavailable.

We set the sample size to $n = 400$ and vary the ambient dimension $p \in \{10, 20, 30\}$. The informative nonlinear structure is embedded in the first three coordinates, while the remaining $p - 3$ variables are independent standard Gaussian noise. We consider four data generating processes featuring complex manifolds and symmetric configurations:

- **Model I:** $X_1 = \theta\cos\theta + \delta_1, X_2 = \theta\sin\theta + \delta_2, X_3 = h + \delta_3$, where $\theta \sim \mathcal{U}(1.5\pi, 4.5\pi)$ and $h \sim \mathcal{U}(-2, 2)$. The label is $Y = \text{sign}(\sin(2.5\theta) + 0.1\epsilon)$.
- **Model II:** $X_1 = (R + r\cos\theta)\cos\phi + \delta_1, \quad X_2 = (R + r\cos\theta)\sin\phi + \delta_2, \quad X_3 = r\sin\theta + \delta_3$, where $\theta, \phi \sim \mathcal{U}(0, 2\pi)$, $R = 6$, and $r = 2$. The label is defined by $Y = \mathbf{1}\{\cos\phi < 0\} \cdot 2 + \mathbf{1}\{\cos\theta < 0\}$, so that $Y \in \{0, 1, 2, 3\}$.
- **Model III:** $X_1 = t\cos t + \delta_1, X_2 = t\sin t + \delta_2, X_3 = t/2 + \delta_3$, where $t \sim \mathcal{U}(0.5, 3\pi)$. The label is $Y = \text{sign}(\sin(2.5t) + 0.1\epsilon)$.

- **Model IV:** We generate 8 cluster centers $\boldsymbol{\mu}$ at the vertices of a 3D hypercube $\{\pm 2.0\}^3$. $X_i = \mu_i + \delta_i$ for $i = 1, 2, 3$. $Y = \text{sign}(X_1 X_2 X_3)$.

Over 100 replications (70% training, 30% testing), we fix the structural dimension at $r = 3$ and evaluate performance using a 5-nearest-neighbor classifier (Table 1). KPOTD consistently achieves the highest classification accuracy across all models. Notably, it exhibits strong robustness to nuisance variables, maintaining accuracies above 0.91 even as p increases to 30. In contrast, inverse-regression methods struggle with structural symmetry, and other nonlinear baselines show significant performance degradation in higher dimensions. These results confirm that KPOTD effectively captures high-order geometric discrepancies.

5.2 Real Data

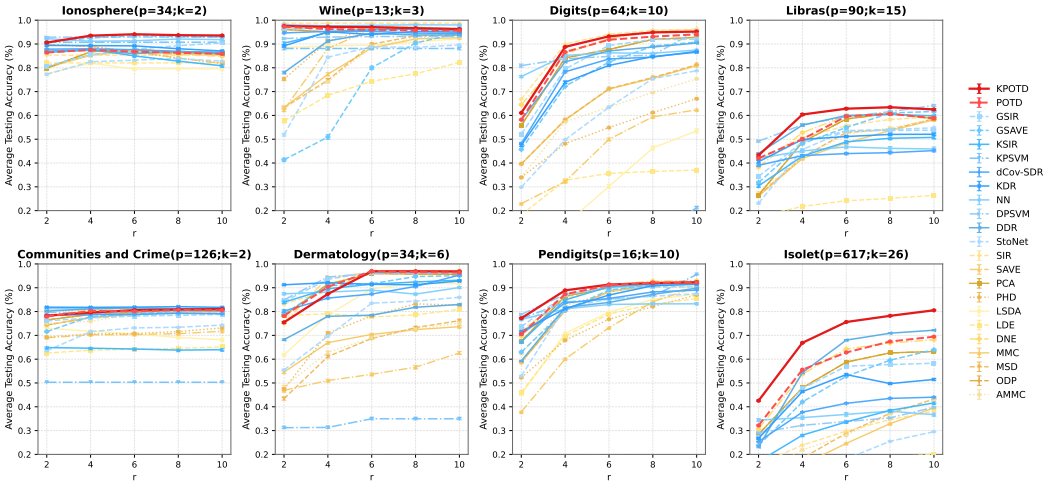

Figure 2: Classification accuracy as a function of the extracted dimension r on eight real datasets. The evaluated methods are color-coded by their respective algorithmic families for clarity: red lines denote optimal-transport-based methods (including the proposed KPOTD), blue lines represent competing nonlinear dimension reduction methods, and yellow lines indicate traditional linear methods. Results are averaged over 100 random train–test splits.

We compare KPOTD against competing methods on eight classification datasets from the UCI Machine Learning Repository: *Ionosphere*, *Wine*, *Digits*, *Libras*, *Communities and Crime*, *Dermatology*, *Pendigits*, and *Isolet*. These datasets cover binary and multi-class tasks ($p \in [13, 617]$, $k \in [2, 26]$). For *Communities and Crime*, the continuous response is dichotomized at the sample median. To maintain computational efficiency, *Isolet* and *Communities and Crime* are stratified subsampled to 1,000 observations. We run 100 replications for each dataset, randomly holding out one-third of the observations for testing.

Baselines include traditional linear methods (PCA, SIR [Li, 1991], SAVE [Cook and Weisberg, 1991], PHD [Li, 1992], AMMC [Lu and Tan, 2011], DNE [Zhang et al., 2006], LDE [Chen et al., 2005], LSDA [Cai et al., 2007], MMC [Li et al., 2006], MSD [Song et al., 2007], ODP [Li et al., 2009]) alongside the nonlinear models from Section 5.1. We evaluate five structural dimensions, $r \in \{2, 4, 6, 8, 10\}$, using a 10-nearest-neighbor classifier trained on the projected data.

As shown in Figure 2, traditional linear methods consistently underperform due to intrinsic nonlinear manifolds. While competing nonlinear baselines fluctuate across datasets, KPOTD consistently achieves top performance, particularly in complex multi-class tasks. This confirms that extracting functional optimal transport displacements yields highly discriminative representations. A detailed execution time analysis is provided in Appendix D.2.

6 Conclusion

This paper introduces Kernel Principal Optimal Transport Directions (KPOTD) for nonlinear sufficient dimension reduction. By integrating optimal transport with functional spaces, KPOTD overcomes the inherent limitations of moment-matching and margin-based paradigms. Computationally, it reduces infinite-dimensional functional optimization to a stable, finite-dimensional generalized eigenvalue problem. Theoretically, we establish the unbiasedness and exhaustiveness of the KPOTD operator, guaranteeing exact central subspace recovery under mild conditions. Extensive experiments confirm that KPOTD consistently outperforms modern baselines, exhibiting exceptional robustness against high-dimensional noise and successfully transforming entangled nonlinear manifolds into linearly separable representations.

Acknowledgments

We thank the anonymous reviewers and the STAI-X 2026 Program Chairs for their careful reading and constructive comments, which helped improve the clarity and presentation of this work. We also thank the members of the BDAL-RUC research group for helpful discussions.

References

Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, and Hujun Bao. Locality sensitive discriminant analysis. In *IJCAI*, volume 2007, pages 1713–1726, 2007.

Hwann-Tzong Chen, Huang-Wei Chang, and Tyng-Luh Liu. Local discriminant embedding and its variants. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 846–853. IEEE, 2005. doi: 10.1109/CVPR.2005.216.

Xin Chen, Changliang Zou, and R. Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723, 2010. doi: 10.1214/10-AOS826.

Yinfeng Chen, Jin Liu, and Rui Qiu. Deep principal support vector machines for nonlinear sufficient dimension reduction. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

R Dennis Cook. *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons, 1998.

R Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. doi: 10.1111/j.1467-9868.2008.00674.x.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

Yukuan Hu, Mengyu Li, Xin Liu, and Cheng Meng. Sampling-based methods for multi-block optimization problems over transport polytopes. *Mathematics of Computation*, 94(353):1281–1322, 2025.

Jian Huang, Yuling Jiao, Xu Liao, Jin Liu, and Zhou Yu. Deep dimension reduction for supervised representation learning. *IEEE Transactions on Information Theory*, 70(5):3583–3598, 2024.

Shayan Hundrieser, Thomas Staudt, and Axel Munk. Empirical optimal transport between different measures adapts to lower complexity. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):824–846, 2024. doi: 10.1214/23-AIHP1369.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2), 2021.

Daniel Kapla, Lukas Fertl, and Efstathia Bura. Fusing sufficient dimension reduction with neural networks. *Computational Statistics & Data Analysis*, 168:107389, 2022.

Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Annals of Statistics*, 41(1):221–249, 2013. doi: 10.1214/12-AOS1071.

Bing Li, Andreas Artemiou, and Lexin Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210, 2011.

Bo Li, Chunheng Wang, and De-Shuang Huang. Supervised feature extraction based on orthogonal discriminant projection. *Neurocomputing*, 73(1-3):191–196, 2009.

Haifeng Li, Tao Jiang, and Keshu Zhang. Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks*, 17(1):157–165, 2006.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420): 1025–1039, 1992.

Lexin Li, Xuewen Wen, and Liping Zhu. A selective overview of sparse sufficient dimension reduction. *Statistical Theory and Related Fields*, 4(2):121–145, 2020. doi: 10.1080/24754269.2020.1829389.

Mengyu Li, Jun Yu, Tao Li, and Cheng Meng. Importance sparsification for Sinkhorn algorithm. *Journal of Machine Learning Research*, 24(247):1–44, 2023a.

Mengyu Li, Jun Yu, Hongteng Xu, and Cheng Meng. Efficient approximation of Gromov-Wasserstein distance using importance sparsification. *Journal of Computational and Graphical Statistics*, 32(4):1512–1523, 2023b. doi: 10.1080/10618600.2023.2165500.

Qian Li, Zhichao Wang, Gang Li, Jun Pang, and Guandong Xu. Hilbert Sinkhorn divergence for optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2021.

Tao Li, Jun Yu, and Cheng Meng. Scalable model-free feature screening via Sliced-Wasserstein dependency. *Journal of Computational and Graphical Statistics*, 32(4):1501–1511, 2023c.

Tao Li, Cheng Meng, Hongteng Xu, and Jun Yu. Hilbert curve projection distance for distribution comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4993–5007, 2024.

Tao Li, Cheng Meng, Hongteng Xu, and Jun Zhu. Efficient variants of Wasserstein distance in hyperbolic space via space-filling curve projection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2025. doi: 10.1109/TNNLS.2025.3551275.

Siqi Liang, Yan Sun, and Faming Liang. Nonlinear sufficient dimension reduction with a stochastic neural network. *Advances in Neural Information Processing Systems*, 35:27360–27373, 2022.

Jiwen Lu and Yap-Peng Tan. Adaptive maximum margin criterion for image classification. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011.

Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, 32, 2019.

Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, and Wenxuan Zhong. Sufficient dimension reduction for classification using principal optimal transport direction. *Advances in Neural Information Processing Systems*, 33:4015–4028, 2020.

Hà Quang Minh. Finite sample approximations of exact and entropic wasserstein distances between covariance operators and gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):96–124, 2022.

Hà Quang Minh. Convergence and finite sample approximations of entropic regularized wasserstein distances in gaussian and rkhs settings. *Analysis and Applications*, 21(03):719, 2023.

Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P Apte, Joseph O Deasy, and Allen Tannenbaum. A novel kernel Wasserstein distance on Gaussian measures: An application of identifying dental artifacts in head and neck computed tomography. *Computers in Biology and Medicine*, 120:103731, 2020.

Xiaxue Ouyang, Hao Zheng, Haoxian Liang, Jingyi Zhang, Yixuan Qiu, Cheng Meng, and Mengyu Li. Sparsification techniques for large-scale optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 18(1):e70056, 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Wenyu Sheng and Xiangrong Yin. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 22(4):1046–1060, 2013.

Fengxi Song, David Zhang, Qiying Chen, and Jingyu Wang. Face recognition based on a novel linear discriminant criterion. *Pattern Analysis and Applications*, 10(3):165–174, 2007.

Kean Ming Tan, Zhaoran Wang, Tong Zhang, Han Liu, and R. Dennis Cook. A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, 105(4):769–782, 2018. doi: 10.1093/biomet/asy049.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2009.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Han-Ming Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

Jingyi Zhang, Wenxuan Zhong, and Ping Ma. A review on modern computational optimal transport methods with applications in biomedical research. *Modern Statistical Methods for Health Research*, pages 279–300, 2021.

Jingyi Zhang, Ping Ma, Wenxuan Zhong, and Cheng Meng. Projection-based techniques for high-dimensional optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1587, 2023.

Wei Zhang, Xiangyang Xue, Hong Lu, and Yuan-Fang Guo. Discriminant neighborhood embedding for classification. *Pattern Recognition*, 39(11):2240–2243, 2006.

Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel Hilbert spaces: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1741–1754, 2020.

Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011. doi: 10.1198/jasa.2011.tm10563.