Variable Selection for Kernel Two-Sample Tests

Jie Wang
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, jwang3163@gatech.edu

Santanu S. Dey
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, santanu.dey@isye.gatech.edu

Yao Xie
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, yao.xie@isye.gatech.edu

We consider the variable selection problem for two-sample tests, aiming to select the most informative variables to distinguish samples from two groups. To solve this problem, we propose a framework based on the kernel maximum mean discrepancy (MMD). Our approach seeks a group of variables with a pre-specified size that maximizes the variance-regularized MMD statistics. This formulation also corresponds to the minimization of asymptotic type-II error while controlling type-I error, as studied in the literature. We present mixed-integer programming formulations and develop exact and approximation algorithms with performance guarantees for different choices of kernel functions. Furthermore, we provide a statistical testing power analysis of our proposed framework. Experiment results on synthetic and real datasets demonstrate the superior performance of our approach*.

1. Introduction

We study the variable selection problem for two-sample testing, which aims to select the most informative variables to distinguish differences in the distributions between two groups of samples. On the one hand, it is crucial to identify interpretable variables that contribute to the inherent differences between populations, as they play a critical role in various scientific discovery areas. For instance, in the context of gene expressions and biological indicators, only a small subset of variables may account for the disparities between normal and abnormal data samples [63]. On the other hand, the dissimilarities between high-dimensional datasets often exhibit a low-dimensional structure [59]. Consequently, extracting a small set of crucial variables as a pre-processing step enhances the efficacy of high-dimensional two-sample testing. The selection of key variables can be challenging primarily for the following reasons: (i) Limited information is available regarding the data distribution for each group; (ii) The number of observed samples is insufficient to obtain accurate estimates of the distributions for each group; (iii) The high dimensionality of the data points makes it challenging to compare the two groups effectively; (iv) The task of variable selection problem is typically formulated as a combinatorial optimization, which is NP-hard to solve in most cases.

The problem of non-parametric variable selection for two-sample testing with limited data samples has been a long-standing challenge in literature. Classical approaches mainly rely on parametric assumptions regarding the data-generating distributions. For example, Taguchi and Rajesh [56] assume target distributions as Gaussian and find important variables such that the difference between mean and covariance among two groups is maximized. Following this seminal work, references [30, 29, 28] further model distributions as Gaussian graphical models and detect the difference between distributions in correlation and partial correlation. However, it is undesirable to restrict the analysis to parametric distributions because those assumptions may not hold for real-world data. The Bonferroni method [10] has been proposed in the two-sample testing context to compare every single feature using statistical tests to obtain representative variables, but it may not perform well when correlations exist between them.

* Our numerical implementation code is online available at https://github.com/WalterBabyRudin/MMDVar_Selection

Recently, Mueller and Jaakkola [47] proposed the projected Wasserstein distance for this task, operating by finding the sparse projection direction such that the univariate Wasserstein distance between projected samples is maximized. Since the Wasserstein distance is flexible enough to compare arbitrarily two distributions even with non-overlapping support [23], this approach serves as a non-parametric way for selecting variables. However, it is important to note that the non-asymptotic convergence of the proposed projected Wasserstein distance depends significantly on factors such as the size of the distribution support and the projected dimension. Furthermore, due to the nonconvex nature of the problem, only approximation algorithms have been proposed to find local optimal solutions. Unfortunately, there is currently no theoretical result available to quantify the sub-optimality gap of the estimated solution. In statistical analysis, the assumption typically made is that one can successfully identify the global optimum, which is not guaranteed in this case.

In addition to Wasserstein distance, the MMD statistics are also popular in signal processing and machine learning areas, motivated by their computational efficiency and nice statistical properties [27, 26, 25]. The MMD-based approaches have been proposed in the literature to study the problems of one- or two-sample testing [27, 26, 25, 40, 31, 15, 53, 54]. In this paper, we leverage the MMD framework to propose a novel approach for variable selection in two-sample testing. Specifically, we aim to select key variables that maximize the variance-regularized MMD statistic, which in turn (approximately) maximizes the corresponding kernel testing power. Our contributions are summarized as follows:

(I) We first provide computation algorithms for an *inhomogeneous* quadratic maximization problem with ℓ_2 and ℓ_0 norm constraints (see Section 3), called **S**parse **T**rust **R**egion **S**ubproblem (STRS), which plays a key role for MMD optimization. Despite NP-hardness, we provide an exact mixed-integer semi-definite programming formulation together with exact and approximation algorithms for solving this problem. To the best of our knowledge, this study is new in the literature.

(II) From the computational aspect, we reformulate the MMD optimization framework for the linear kernel case as STRS (see Section 4.1), which can be solved based on our proposed algorithms. For generic kernel cases, the MMD optimization becomes a sparse maximization of a non-concave function (see Section 4.2), which is intractable in general. We propose a heuristic algorithm that iteratively optimizes a quadratic approximation of the objective function, which is also a special case of STRS.

(III) From the statistical aspect, we first provide a convergence analysis regarding our proposed objective function. Next, we demonstrate the consistency of testing power and the rate of type-II risk of our proposed framework (see Section 5).

(IV) Finally, we conduct numerical experiments with synthetic and real datasets to demonstrate the superior performance of our proposed framework over other baseline models.

Notations. Given a positive integer n, define $[n] = \{1, \ldots, n\}$. Let $\mathbb{F} = \{0, 1\}$, and $\mathbb{S}_n^+$ denote the collection of $n \times n$ symmetric positive semi-definite matrices. Given a vector $z \in \mathbb{R}^D$ and a set $S \subseteq [D]$, we use $z^{(k)}$ denote the k-th entry in z, and $z^{(S)}$ to denote the subvector with entries indexed by S. Given an $m \times n$ matrix A and two sets $S \subseteq [m], t \subseteq [n]$, denote $A^{(i,j)}$ the (i,j)-th entry in A and denote $A^{(S,T)}$ as the submatrix with rows and columns indexed by S and T. Given a vector $z \in \mathbb{R}^D$ and a distribution μ in $\mathbb{R}^D$, denote $z \circ \mu$ as the distribution of the random variable $\sum_{k\in[D]} z^{(k)} x^{(k)}$ provided that $x \sim \mu$. Define the norm $\|z\|_{(d)} = \max_{S:\ |S|\le d} \left\|z^{(S)}\right\|_2$.

1.1. Related Work

Variable selection. Classical variable selection approaches seek to extract the most valuable features from a group of high-dimensional data points. In particular, the sparse PCA approach seeks to select crucial variables that maximize the sample covariance based on sample sets [37, 18, 17]; the truncated SVD approach aims to formulate a low-rank data matrix with minimum approximation error [36], and the maximum entropy sampling or experiment design approach aims to select a subgroup of

samples that reserve information as much as possible [38, 35]. However, the literature has paid less attention to variable selection for identifying differences between the two groups. Recently, Mueller and Jaakkola [47] proposed to find the optimal subset of features such that the Wasserstein distance between projected distributions in dimension $d = 1$ is maximized. Later Wang et al. [60, 61] modified the projection function as the linear mapping with general dimension $d > 1$ and nonlinear mapping, respectively, thus improving the flexibility of dimensionality reduction and power of two-sample testing. Nevertheless, these references do not impose sparsity constraints when performing dimensionality reduction, and therefore, they are unable to select a subset of variables that differentiate the differences between the two groups.

Kernel-based two-sample tests. A popular approach for non-parametric two-sample testing is based on kernel methods [49]: such tests quantify the difference of probability distributions by measuring the difference in *kernel mean embeddings* [6, 46], which is also called the maximum mean discrepancy (MMD) in the lierature [25, 21, 32, 50, 51]. The follow-up references [40, 55] further improve the performance of kernel-based two-sample tests by selecting kernels that maximize the variance-normalized empirical MMD. We adopt this idea in our variable selection framework. However, we observe that using this criterion for variable selection results in a fractional program subject to sparsity and norm constraints, which is highly challenging to solve. Hence, we are inspired to consider optimizing the variance-regularized empirical MMD statistic as a surrogate.

Classifier-based two-sample tests. Some widely-used hypothesis testing frameworks employ classification techniques for two-sample testing (see, e.g., [14, 34, 33]). It is worth noting that our approach adopts a distinct framework compared to those references: these aforementioned testing methods may not effectively identify interpretable variables capable of distinguishing between two distributions. One potential alternative is to employ a classifier based on sparse logistic regression [8] to construct a two-sample test. However, this approach may not yield satisfactory performance due to the limited flexibility of the parametric form of the classifier, as we will demonstrate in Section 2.1.

2. Model Formulation

Let $\mathbf{x}^n := \{x_i\}_{i=1}^n$ and $\mathbf{y}^n = \{y_i\}_{i=1}^n$ be n i.i.d. samples generated from distributions μ and ν, respectively. We assume the sample sizes from these two groups are equal for notational simplicity, but our results can be easily extended for cases with unequal sample sizes. In particular, those data samples are in the Euclidean space $\mathbb{R}^D$, where the dimension D denotes the number of feature variables. In the following, we present a variable selection framework for two-sample testing, aimed at identifying the most informative variables that can distinguish whether the distributions μ and ν are different.

We first present some background information about maximum mean discrepancy (MMD). MMD measures the discrepancy between two probability distributions by employing test functions within a reproducing kernel Hilbert space (RKHS), which has been commonly used in two-sample testing literature [27, 26, 25, 40, 31, 15].

Definition 1 (Maximum Mean Discrepancy). A kernel function $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is called a positive semi-definite kernel if for any finite set of n samples $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$ and $\{c_i\}_{i=1}^n$ in $\mathbb{R}$, it holds that $\sum_{i\in[n]}\sum_{j\in[n]} c_i c_j K(x_i, x_j) \geq 0$. A positive semi-definite kernel K induces a unique RKHS $\mathcal{H}$. Given a RKHS $\mathcal{H}$ containing a class of candidate testing functions and two distributions μ, ν, define the corresponding MMD statistic as

$$\mathrm{MMD}(\mu, \nu; K) \triangleq \sup_{f\in\mathcal{H}, \|f\|_{\mathcal{H}}\leq 1} \Big\{ \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] \Big\}. \qquad \diamondsuit$$

Leveraging reproducing properties of the RKHS, the MMD statistic can be equivalently written as

$$\mathrm{MMD}^2(\mu, \nu; K) = \mathbb{E}_{x,x'\sim\mu}[K(x, x')] + \mathbb{E}_{y,y'\sim\nu}[K(y, y')] - 2\mathbb{E}_{x\sim\mu, y\sim\nu}[K(x, y)],$$

which enables convenient computation and sample estimation. When the distributions μ and ν are not available, one can formulate an estimate of $\mathrm{MMD}^2(\mu,\nu;K)$ based on samples $\mathbf{x}^n$ and $\mathbf{y}^n$ using the following statistic [25]:

$$\widehat{\mathrm{MMD}}^2(\mathbf{x}^n,\mathbf{y}^n;K) \triangleq \frac{1}{n(n-1)} \sum_{i\in[n],j\in[n],i\neq j} H_{i,j}, \tag{1}$$

with

$$H_{i,j} := K(x_i,x_j)+K(y_i,y_j)-K(x_i,y_j)-K(y_i,x_j). \tag{2}$$

The choice of kernel function largely influences the performance of variable selection for two-sample tests. To achieve satisfactory performance, we consider the following types of kernel functions, denoted as $K_z(\cdot,\cdot)$. Here, the coefficient vector $z=(z^{(s)})_{s\in[D]}$ involved in the kernel functions determines which variables to be selected, which is in the domain set

$$\mathcal{Z} := \{z\in\mathbb{R}^D : \|z\|_2=1, \|z\|_0\le d\}. \tag{3}$$

- **Linear Kernel**: For each coordinate $s\in[D]$, we specify the scalar-input kernel $k_s:\ \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and then construct
$$K_z(x,y)=\sum_{s\in[D]} z^{(s)}k_s\big(x^{(s)},y^{(s)}\big). \tag{4}$$
Those scalar-input kernels $k_s(\cdot,\cdot), s\in[D]$ defined above are used to compare the difference of distributions among each coordinate, which can be chosen as the Gaussian kernel with certain bandwidth hyper-parameter τ_s^2, i.e., $k_s(x,y)=e^{-(x-y)^2/(2\tau_s^2)}$.
- **Quadratic Kernel**: For each coordinate $s\in[D]$, we specify the scalar-input kernel $k_s:\ \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and then construct
$$K_z(x,y)=\left(\sum_{s\in[D]} z^{(s)}k_s\big(x^{(s)},y^{(s)}\big)+c\right)^2. \tag{5}$$
Here $c\ge 0$ is a bandwidth hyper-parameter of the quadratic kernel, and scalar-input kernels $k_s(\cdot,\cdot), s\in[D]$ can be chosen in the same way as defined in the linear kernel case.
- **Gaussian Kernel**: We first specify the bandwidth hyper-parameter $\sigma^2>0$ and then construct
$$K_z(x,y)=\exp\left(-\frac{\sum_{s\in[D]}\big(z^{(s)}(x^{(s)}-y^{(s)})\big)^2}{2\sigma^2}\right). \tag{6}$$

We pick the sparse selection vector z to achieve the most powerful test. Inspired by the fact that the kernel function leading to the most powerful two-sample test approximately maximizes the MMD testing statistic normalized by its variance [55], we aim to pick the selection vector z that solves the following optimization problem:

$$\max_{z\in\mathcal{Z}} \quad \left\{\widehat{F}(z;\mathbf{x}^n,\mathbf{y}^n) := \widehat{\mathrm{MMD}}^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \lambda\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)\right\}, \tag{7}$$

where $\lambda>0$ is a regularization hyper-parameter that can be tuned by cross-validation. Here $\widehat{\mathrm{MMD}}^2(\mathbf{x}^n,\mathbf{y}^n;K_z)$ and $\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)$ are unbiased empirical estimators of the population testing statistic and the variance of testing statistic under alternative hypothesis $\mathcal{H}_1:\ \mu\neq\nu$, respectively. For fixed samples $\mathbf{x}^n,\mathbf{y}^n$ and kernel function $K(\cdot,\cdot)$, by [55], the variance estimator

$$\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K)=\frac{4}{n^3}\sum_{i\in[n]}\left(\sum_{j\in[n]}H_{i,j}\right)^2-\frac{4}{n^4}\left(\sum_{i\in[n]}\sum_{j\in[n]}H_{i,j}\right)^2, \tag{8}$$

Algorithm 1 A permutation two-sample test using MMD with variable selection

Require: cardinality d, type-I error threshold α_{level}, bootstrap size N_p, collected samples $\mathbf{x}^n$ and $\mathbf{y}^n$.

1: Split data as $\mathbf{x}^n = \mathbf{x}^{\text{Tr}} \cup \mathbf{x}^{\text{Te}}$ and $\mathbf{y}^n = \mathbf{y}^{\text{Tr}} \cup \mathbf{y}^{\text{Te}}$.
2: Solve (7) with input data $(\mathbf{x}^{\text{Tr}}, \mathbf{y}^{\text{Tr}})$ to obtain optimal sparse selection vector z^*.
3: Compute test statistic $T = \widehat{\text{MMD}}^2(\mathbf{x}^{\text{Te}}, \mathbf{y}^{\text{Te}}; K_{z^*})$.
4: {decide threshold}
5: **for** $i = 1, \ldots, N_p$ **do**
6: Shuffle $\mathbf{x}^{\text{Te}} \cup \mathbf{y}^{\text{Te}}$ to obtain $\mathbf{x}^{\text{Te}}_{(i)}$ and $\mathbf{y}^{\text{Te}}_{(i)}$.
7: Compute test statistic for bootstrap samples $T_i = \widehat{\text{MMD}}^2(\mathbf{x}^{\text{Te}}_{(i)}, \mathbf{y}^{\text{Te}}_{(i)}; K_{z^*})$.
8: **end for**
9: $t_{\text{thres}} \leftarrow (1 - \alpha_{\text{level}})$-quantile of $\{T_i\}_{i \in [N_p]}$.
10: Reject $\mathcal{H}_0$ (i.e., decide the two sample distributions are different) if $T > t_{\text{thres}}$.

where $H_{i,j}, i, j \in [n]$ are defined in (2). The rationale behind problem (7) is that, by properly tuning the regularizer $\lambda > 0$, we balance the trade-off between maximizing the testing statistic and minimizing its variance, which amounts to approximately optimize the testing power criteria.

Using the proposed variable selection framework, we present a kernel two-sample test as follows. The data points are divided into training and testing datasets. Initially, the training set is utilized to obtain the selection coefficient that optimally identifies the differences between the two groups. Next, a permutation test is performed on the testing data points, projected based on the trained selection coefficient. The detailed algorithm is presented in Algorithm 1. This test is guaranteed to control the type-I error [24] because we evaluate the p-value of the test via the permutation approach. In the following sections, we discuss how to solve the optimization problem (7) with linear and quadratic kernels, respectively. In the subsequent sections, we first provide optimization algorithms for solving a special mixed-integer quadratic programming (MIQP), which plays a key role in MMD optimization. Next, we develop tractable algorithms for solving the MMD optimization. Finally, we establish statistical testing power guarantees for our proposed framework.

2.1. Connections with Classification-Based Testing

It is worth noting that our proposed method can be viewed as a generalized classifier-based testing using logistic loss. Specifically, our proposed method has the following two phases:

(I) At the first phase, we choose a suitable kernel function $K(\cdot, \cdot)$ based on training data $\mathbf{x}^{\text{Tr}}$ and $\mathbf{y}^{\text{Tr}}$ that depends only on a small group of variables leading to satisfactory two-sample testing performance. Such a variable selection procedure makes our classification model more interpretable.

(II) At the second phase, we obtain the witness function (see, e.g., [25, Section 2.3]), denoted as $\widehat{f}$, based on validation data $\mathbf{x}^{\text{Te}}$ and $\mathbf{y}^{\text{Te}}$:

$$\widehat{f}(z) \propto \frac{1}{|\mathbf{x}^{\text{Te}}|} \sum_{x \in \mathbf{x}^{\text{Te}}} K(x, z) - \frac{1}{|\mathbf{y}^{\text{Te}}|} \sum_{y \in \mathbf{y}^{\text{Te}}} K(y, z). \tag{9}$$

Consequently, for a new given sample z, we classify it into group X with data distribution μ or group Y with data distribution ν using the Bayesian rule

$$\mathbb{P}(z \in X) = \frac{e^{\widehat{f}(z)}}{1 + e^{\widehat{f}(z)}}, \qquad \mathbb{P}(z \in Y) = \frac{1}{1 + e^{\widehat{f}(z)}}$$

A notable existing variable selection approach for classification is the *sparse logistic regression* (SLR) [8], aiming to classify z according to the rule

$$\mathbb{P}(z \in X) = \frac{e^{\beta^{\mathrm{T}} z}}{1+e^{\beta^{\mathrm{T}} z}}, \quad \mathbb{P}(z \in Y) = \frac{1}{1+e^{\beta^{\mathrm{T}} z}}$$

for some sparse vector β. The non-zero entries of the coefficient vector β correspond to the selected variables that distinguish the differences between groups X and Y. This approach can also be used for two-sample testing: based on samples $\mathbf{x}^{\mathrm{Te}}$ and $\mathbf{y}^{\mathrm{Te}}$, we formulate the following testing statistic and reject the null hypothesis if it exceeds a certain threshold:

$$T_{\mathrm{SLR}} = \frac{1}{|\mathbf{x}^{\mathrm{Te}}|} \sum_{x \in \mathbf{x}^{\mathrm{Te}}} \beta^{\mathrm{T}} x - \frac{1}{|\mathbf{y}^{\mathrm{Te}}|} \sum_{y \in \mathbf{y}^{\mathrm{Te}}} \beta^{\mathrm{T}} y.$$

Such an approach assumes a parametric assumption that the data distributions μ and ν are *linearly separable* since otherwise the linear predictor may not achieve satisfactory performance. In contrast, our proposed approach constructs a non-parametric and kernel-based classifier defined in (9), which seems more flexible for real-world applications.

In the following, we provide an example to demonstrate that our proposed framework can successfully select useful variables to distinguish the difference between two groups, while the sparse logistic regression cannot finish this task.

Example 1 (Example when sparse logistic regression cannot identify variables). Consider the example where $\mu = \mathcal{N}(0, I_D)$ and $\nu = \mathcal{N}(0, \mathrm{diag}((1+\epsilon)^2, 1, \ldots, 1))$ with $\epsilon > 0$. Here, only the first coordinate can differentiate between μ and ν. When using the sparse logistic regression, it is clear that for any β satisfying $\|\beta\|_0 \leq 1$, it holds that the population version of testing statistic $\mathbb{E}[T_{\mathrm{SLR}}] = 0$. This indicates that sparse logistic regression may not achieve satisfactory performance in hypothesis testing or classification. In contrast, consider our proposed MMD framework with the linear kernel. For any z such that $\|z\|_2 = 1, \|z\|_0 \leq 1$, it holds that the population version of the objective in (7) achieves the unique optimal solution $\widehat{z}$ with $\widehat{z}^{(1)} = 1$ if the variance regularization λ is selected properly. Specifically, when λ is chosen to be smaller than a constant $\bar{\lambda} > 0$, our proposed MMD framework can always select the true useful variable.[1] ♣

3. Sparse Trust Region Subproblem (STRS)

To achieve variable selection in our setting, we need to solve the following mixed-integer quadratic program (MIQP):

$$\max_{z \in \mathcal{Z}} \left\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} t \right\}, \tag{STRS}$$

where the set $\mathcal{Z}$ is defined in (3) and (A, t) are input coefficients to be specified later. Without loss of generality, we assume $A \succeq 0$, since otherwise, we can re-write the problem as

$$\max_{z \in \mathcal{Z}} \left\{ z^{\mathrm{T}} (A - \lambda_{\min}(A) I_D) z + z^{\mathrm{T}} t \right\} + \lambda_{\min}(A),$$

where the shifted matrix $A - \lambda_{\min}(A) I_D \succeq 0$, where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix. It is worth mentioning that the problem (STRS) reduces to sparse PCA formulation when the coefficient vector $t = 0$, which has been studied extensively in the literature [5, 22, 45, 37]. However, the study for general vector t for the problem (STRS) is new. In the following, we discuss the exact and approximation algorithms for solving (STRS) with generic data matrix A and vector t.

There are two challenges solving (STRS) in particular for large-scale problems. First, since the objective function is non-concave in z, it is difficult to develop exact algorithms directly for solving (STRS). Instead, we provide a mixed-integer *convex* programming reformulation, which motivates us to develop exact algorithms in Section 3.1. Second, this problem is NP-hard even if $t = 0$, as pointed out in [43]. When the problem is large-scale, we provide approximation algorithms with provable performance guarantees.

3.1. Exact Mixed-Integer SDP (MISDP) Reformulation

We first provide an exact MISDP reformulation of (STRS). When the coefficient vector $t = 0$, similar reformulation results have been developed in the sparse PCA literature [37, 7]. However, such a reformulation for $t \neq 0$ is new in the literature. For notational simplicity, we define the following block matrix of size $(D+1) \times (D+1)$:

$$\tilde{A} = \begin{pmatrix} 0 & \frac{1}{2}t^{\mathrm{T}} \\ \frac{1}{2}t & A \end{pmatrix}.$$

THEOREM 1 **(MISDP Reformulation of (STRS)).** *Problem (STRS) can be equivalently formulated as the following MISDP*

$$\max_{Z \in \mathbb{S}^+_{D+1}, q \in \mathcal{Q}} \quad \langle \tilde{A}, Z \rangle \tag{10a}$$

$$\text{s.t.} \quad Z^{(i,i)} \leq q^{(i)}, \quad i \in [D], \tag{10b}$$

$$Z^{(0,0)} = 1, Tr(Z) = 2, \tag{10c}$$

where the set

$$\mathcal{Q} = \left\{ q \in \mathbb{F}^D : \sum_{k \in [D]} q^{(k)} \leq d \right\}, \tag{11}$$

and we assume the indices of $Z, \tilde{A} \in \mathbb{S}^+_{D+1}$ are both over $[0:D] \times [0:D]$. The continuous relaxation value of (10) equals $w_{\mathrm{rel}} = \max_{z:\ \|z\|_2 = 1} \{z^{\mathrm{T}} A z + z^{\mathrm{T}} t\}$.

The proof idea of Theorem 1 is to express the problem (STRS) as a *rank*-1 *constrained SDP* problem. Leveraging well-known results on rank-constrained optimization (see, e.g., [48, 16, 39]), one can remove the rank constraint without changing the optimal value of the original SDP problem. Although (10) is equivalent to (STRS), the fact that its continuous relaxation value is equal to w_{rel} suggests that it may be a weak formulation. Inspired from [37, 7], we propose the additional two valid inequalities to strengthen the formulation (10) in Corollay 1.

COROLLARY 1 **(Stronger MISDP Reformulation of (STRS)).** *The problem (STRS) reduces to the following stronger MISDP formulation:*

$$\max_{Z \in \mathbb{S}^+_{D+1}, q \in \mathcal{Q}} \quad \langle \tilde{A}, Z \rangle \tag{12a}$$

$$\text{s.t.} \quad (10c), \sum_{j \in [D]} (Z^{(i,j)})^2 \leq Z^{(i,i)} q^{(i)}, \left(\sum_{j \in [D]} |Z^{(i,j)}| \right)^2 \leq d Z^{(i,i)} q^{(i)}, \quad \forall i \in [D]. \tag{12b}$$

It is worth noting that two distinct references [37, 7] have independently introduced two valid inequalities to enhance the performance of solving the sparse PCA problem, which is a special instance of (STRS) for $t = 0$. However, one of the valid inequalities in (12b) proposed in [7] is dominated by a valid inequality proposed in [37], while the other valid inequality has been proposed simultaneously in these two references. This motivates us to incorporate two valid inequalities from [37] into our formulation, as outlined in Corollary 1. On the one hand, the resulting formulation (12) can be directly solved via some exact MISDP solvers such as YALMIP [41]. On the other hand, it enables us to develop a customized exact algorithm to solve this formulation based on Benders decomposition since the binary vector q can be separated from other decision variables.

To develop the exact algorithm, we first reformulate the problem (12) as a max-min saddle point problem so that it can be solved based on the outer approximation technique [ref].

THEOREM 2 **(Saddle Point Reformulation of (12)).** *Problem (12) shares the same optimal value as the following problem:*

$$\max_{q\in\mathcal{Q}} \left\{ f(q) \triangleq \max_{Z\in\mathbb{S}^+_{D+1}} \left\{ \langle \tilde{A}, Z\rangle : s.t.\ (12b) \right\} \right\}. \tag{13}$$

Here the function $f(q)$ is concave in q over the domain $\overline{\mathcal{Q}} := \mathrm{conv}(\mathcal{Q})$, and equivalently, is the optimal value to the following problem:

$$\begin{aligned}
\min_{\lambda,\lambda_0,\nu_1,\nu_2,\Lambda,\beta,\mu,W_1,W_2} \quad & \lambda_0 + 2\lambda + q^{\mathrm{T}}\left[\frac{d}{2}(\nu_1-\nu_2)+\frac{1}{2}(\mu-\mathrm{diag}(\Lambda))\right] \\
& \begin{pmatrix} -\lambda_0 & \frac{1}{2}t^{\mathrm{T}} \\ \frac{1}{2}t & A-\lambda I_D + W_1 - W_2 + \Lambda + \frac{1}{2}\mathrm{diag}(\nu_1+\nu_2) \end{pmatrix} \succeq 0, \quad W_1+W_2-\mathrm{diag}(\beta)\le 0, \\
& \sum_j (\Lambda^{(i,j)})^2 \le (\mu^{(i)})^2, \quad (\beta^{(i)})^2 + (\nu_2^{(i)})^2 \le (\nu_1^{(i)})^2, \quad i\in[D], \\
& \nu_1,\beta,\mu\in\mathbb{R}^D_+, \quad W_1,W_2\in\mathbb{R}^{D\times D}_+, \quad \lambda,\lambda_0\in\mathbb{R}, \quad \nu_2\in\mathbb{R}^D, \Lambda\in\mathbb{R}^{D\times D}.
\end{aligned} \tag{14}$$

For fixed q, the sup-gradient of f with respect to q can be computed as

$$\partial f(q) = \frac{d}{2}(\nu_1^*-\nu_2^*)+\frac{1}{2}(\mu^*-\mathrm{diag}(\Lambda^*)),$$

where $(\nu_1^*,\nu_2^*,\mu^*,\Lambda^*)$ is an optimal solution to the optimization problem above.

By Theorem 2, we find that given a reference direction $\hat{q}$,

$$f(q)\le \bar{f}(q;\hat{q}) \triangleq f(\hat{q}) + g_{\hat{q}}^{\mathrm{T}}(q-\hat{q}),$$

where $g_{\hat{q}}$ is a sup-gradient of f at $\hat{q}$.

Based on this observation, we use the common outer-approximation technique, which is widely used for general mixed-integer nonlinear programs [20, 9], to solve the problem: at iterations $i=1,2,\ldots,i_{\max}-1$, we maximize and refine a piecewise linear upper-bound of $f(q)$:

$$\bar{f}^i(q) = \min_{1\le j\le i} \bar{f}(q;q_j).$$

The algorithm is summarized in Algorithm 2.

By the reference [20], it can be shown that this algorithm yields a non-increasing sequence of overestimators $\{\bar{f}^i(q)\}_{i=1}^{i_{\max}}$, which converge to the optimal value of $f(q)$ within a finite number of iterations $i_{\max}\le \binom{D}{1}+\cdots+\binom{D}{d}$.

Algorithm 2 Exact Algorithm for solving (STRS)

1: **Input:** Max iterations $i_{\max}$, initial guess q_1, tolerance ϵ.
2: **for** $i=1,\ldots,i_{\max}-1$ **do**
3: Compute q_{i+1} as the optimal solution from

$$\max_{q\in\mathcal{Q}} \left\{ \bar{f}^i(q) \triangleq \min_{1\le j\le i} \bar{f}(q;q_j) \right\}$$

4: Compute $f(q_{i+1})$ and $g_{q_{i+1}}\in\partial f(q_{i+1})$
5: **Break** if $f(q_{i+1})-\bar{f}^i(q_{i+1})<\epsilon$
6: **end for**
7: **Return** $q_{i_{\max}}$

3.2. Approximation Algorithms

3.2.1. Convex Relaxation Algorithm. Inspired by Theorem 2, a natural idea of approximately solving the problem (10) is to consider the following problem, in which we replace the nonconvex constraint $q\in\mathcal{Q}$ by a set of linear constraints, which forms its convex hull:

$$\max_{q\in\overline{\mathcal{Q}}} f(q), \quad \text{where } \overline{\mathcal{Q}} = \mathrm{conv}(\mathcal{Q}) = \left\{ q\in[0,1]^D : \sum_i q^{(i)} \le d \right\}. \tag{15}$$

Since the problem (15) is a convex program, it can be solved in polynomial time. Besides, one can obtain a high-quality feasible solution to the problem (10), using a greedy rounding scheme: We first solve (15) to obtain its optimal solution $\tilde{q}$, and then project it onto $\overline{\mathcal{Q}}$ to obtain q. Next, we solve the problem (10) by fixing the variable q and optimizing Z only.

In the following theorem, we provide the approximation ratio regarding the SDP formulation above. The proof adopts similar techniques as in [37, Theorem 5], but we extend the analysis for inhomogeneous quadratic maximization formulation.

THEOREM 3 **(Approximation Gap for Convex Relaxation)**. *Denote by optval(15) and optval(10) the optimal values of problem (15) and (10), respectively. Then, it holds that*

$$\textit{optval}(10) \le \textit{optval}(15) \le \|t\|_2 + \min\left\{ D/d \cdot \textit{optval}(10), d \cdot \textit{optval}(10) - \min_k |t^{(k)}| \right\}.$$

Despite the convexity of problem (15), it is challenging to solve especially for high-dimensional scenarios. References [7, 37] solved a special case of problem (15) when $t = 0$ based on the interior point method (see, e.g., [2, 12, 57]). Unfortunately, since the constraint set of (15) involves the intersection of a semidefinite cone and a large number of second-order cones, re-writing it as a standard conic program and using off-the-shelf solvers to solve this problem spends lots of time. Shiqian [42] designed a novel variable-splitting technique and proposed a first-order Alternating Direction Method of Multipliers [11] (ADMM) algorithm to solve a special convex relaxation of sparse PCA. Unlike this reference that only considers the simplest convex relaxation of sparse PCA without adding strong inequalities, our problem (15) has considerably complicated constraints.

Inspired by the reference [42], we use a similar variable-splitting technique to split the second-order conic constraints and all the other constraints in two blocks of variables, and then propose an ADMM algorithm to optimize the augmented Lagrangian function. The advantage is that each subproblem in iteration update involves only second-order conic constraints or other constraints that are easy to deal with, which results in considerably fast computational speed. We provide a detailed implementation of the proposed algorithm for solving (15) in Appendix EC.1.

3.2.2. Truncation Algorithms with Tighter Approximation Gap. Unfortunately, the SDP relaxation formulation is still challenging to solve for extremely high-dimension scenarios, which motivates us to develop the following computationally cheap truncation approximation algorithms. Compared with the approximation ratio of relaxed SDP formulation in Theorem 3 (i.e., $\min(D/d, d) + \mathcal{O}(1)$), the ratio for our proposed algorithm is tighter (i.e., $\min(D/d, \sqrt{d}) + \mathcal{O}(1)$). We first introduce the definition of normalized sparse truncation operator.

DEFINITION 2 (NORMALIZED SPARSE TRUNCATION). For a vector $z \in \mathbb{R}^D$ and an integer $d \in [D]$, we say $\bar{z}$ is a d-sparse truncation of z if

$$\bar{z}^{(i)} = \begin{cases} z^{(i)}, & \text{if } |z^{(i)}| \text{ is one of the } d \text{ largest (in absolute value) entries in } z \\ 0, & \text{otherwise.} \end{cases}$$

Besides, the vector $\widehat{z} = \bar{z}/\|\bar{z}\|_2$ is said to be the normalized d-sparse truncation of z.

Now, we introduce the following two truncation algorithms:
Truncation Algorithm (I): Let $A^{(:,i)}$ be the i-th column of A for $i \in [D]$, and denote by $\hat{z}_i$ the normalized d-sparse truncation of $A^{(:,i)}$. Then return the estimated optimal solution as the best over all $\hat{z}_i$'s and e_i's for $i \in [D]$, where e_i denotes the i-th standard basis vector.
Truncation Algorithm (II): Relax the ℓ_0-norm constraint in (STRS) and solve the trust region problem $\max_{z:\ \|z\|_2 \le 1} \{z^{\mathrm{T}} A z + z^{\mathrm{T}} t\}$ to obtain the optimal primal solution v. Then, return the estimated optimal solution z as the normalized d-sparse truncation of v.

We summarize the approximation ratios of these two truncation algorithms in Theorem 4. Its proof technique is adopted from [13]. The difference is that the authors consider the approximation ratio under the case $t = 0$, while we adopt the structure of in-homogeneous quadratic function maximization to extend the case for the general coefficient vector t.

THEOREM 4 **(Approximation Gap for Truncation Algorithm).** (I) *Truncation Algorithm* (I) *returns a feasible solution of* (STRS) *with objective value* $V_{(\mathrm{I})}$ *such that*

$$\mathit{optval}(\text{STRS}) \geq V_{(\mathrm{I})} \geq \frac{1}{\sqrt{d}}\mathit{optval}(\text{STRS}) - 2\|t\|_{(d+1)}.$$

(II) *Truncation Algorithm* (II) *returns a feasible solution of* (STRS) *with objective value* $V_{(\mathrm{II})}$ *such that*

$$\mathit{optval}(\text{STRS}) \geq V_{(\mathrm{II})} \geq \frac{d}{D} \cdot \mathit{optval}(\text{STRS}) - \frac{d}{D} \cdot \|t\|_2 - \left(1 + \sqrt{\frac{d}{D}}\right) \cdot \|t\|_{(d)}.$$

We return the best over the output from Truncation Algorithm (I) and (II) as the estimated optimal solution. By Theorem 4, we find the returned solution approximates the optimal solution up to approximation ratio $\min(D/d, \sqrt{d}) + \mathcal{O}(1)$. It has been shown in Chan et al. [13] that it is NP-hard to implement any algorithm with *constant* approximation ratio. Therefore, it is of research interest to explore polynomial-time approximation algorithms with approximation ratio that has milder dependence on D and d. Instead of trying this direction, in the next subsection, we propose another approximation algorithm such that, though NP-hard to solve, it achieves a higher approximation ratio.

3.2.3. Approximation Algorithm via Convex Integer Programming. In this part, we propose an approximation algorithm based on convex integer programming. We first consider the following ℓ_1-norm relaxation of the problem (STRS), which plays a key role in developing our algorithm:

$$\max \left\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} t : \ \|z\|_2 \leq 1, \|z\|_1 \leq \sqrt{d} \right\}. \tag{16}$$

This problem is a relaxation of problem (STRS) because constraints $\|z\|_2 \leq 1, \|z\|_0 \leq d$ imply $\|z\|_1 \leq \sqrt{d}$. Following the similar proof technique as in [17, Theorem 1], we show that solving this new problem results in a constant approximation ratio. The difference is that the authors therein only consider the special case of (STRS) with $t = 0$, while we extend their analysis for general inhomogeneous quadratic objective functions.

THEOREM 5 **(Approximation Gap for ℓ_1-Norm Relaxation).** *There exists a factor* $\rho \in (1, 1 + \sqrt{d/(d+1)}]$ *such that*

$$\mathit{optval}(\text{STRS}) \leq \mathit{optval}(16) \leq \rho^2 \mathit{optval}(\text{STRS}) + (\rho^2 - \rho)\|t\|_2.$$

Although the problem (16) is a relaxation of (STRS), it is still intractable to solve due to the non-concavity of the objective function (recall that $A \succeq 0$). We adopt techniques from [17, Section 2.2] to derive a further convex integer program that serves as a further relaxation of the relaxation problem (16). Before proceeding, we define the following notations. For $i \in [D]$, denote by (λ_i, v_i) the i-th eigen-pair of the matrix A, denote

$$\theta_i := \max\{z^{\mathrm{T}} v_i : \ \|z\|_2 \leq 1, \|z\|_1 \leq \sqrt{d}\},$$

and let $\gamma_i^{[-N:N]}$ be the set of partition points of the domain $[-\theta_i, \theta_i]$, i.e.,

$$\gamma_i^j = \frac{j}{N}\theta_i, \quad j = -N, \ldots, N.$$

Let $\lambda_0 \in \mathbb{R}_+$ be a fixed number such that $\lambda_0 \leq \mathrm{optval}(\text{STRS})$.

PROPOSITION 1 **(Convex Integer Programming Relaxation of (16)).** *Consider the convex integer program:*

$$\textit{Maximize} \quad \lambda_0 + \textstyle\sum_{i:\ \lambda_i > \lambda_0} (\lambda_i - \lambda_0)\xi_i - s \tag{17a}$$

that is subject to the following constraints:

$$\begin{cases} g_i = z^{\mathrm{T}} v_i, \\ |g_i| \le \theta_i, \end{cases} \quad i \in [D], \tag{17b}$$

$$\begin{cases} g_i = \displaystyle\sum_{j \in [-N,N]} \gamma_i^j \eta_i^j, \\ \xi_i = \displaystyle\sum_{j \in [-N,N]} (\gamma_i^j)^2 \eta_i^j, \quad i \in \{i:\ \lambda_i > \lambda_0\}, \\ \eta_i^{[-N,N]} \in \text{SOS-2}, \end{cases} \tag{17c}$$

$$\begin{cases} \displaystyle\sum_{i \in [D]} (z^{(i)})^2 \le 1, \\ \displaystyle\sum_{i:\ \lambda_i > \lambda_0} \left(\xi_i - \frac{\theta_i^2}{4N^2} \right) + \sum_{i:\ \lambda_i \le \lambda_0} g_i^2 \le 1, \end{cases} \tag{17d}$$

$$\begin{cases} \displaystyle\sum_{i \in [D]} y_i \le \sqrt{d}, \\ y_i \ge \left| z^{(i)} \right|, i \in [D], \end{cases} \tag{17e}$$

$$\sum_{i:\ \lambda_i < \lambda_0} -(\lambda_i - \lambda_0) g_i^2 - z^{\mathrm{T}} t \le s, \tag{17f}$$

with SOS-2 *denoting the special ordered set of type-2 [4], and involves the following decision variables:*

$$\{g_i\}_{i=1}^D \in \mathbb{R}^D, \quad \{\xi_i\}_{i \in \{i:\ \lambda_i > \lambda_0\}} \in \mathbb{R}^{|\{i:\ \lambda_i > \lambda_0\}|}, \quad \{\eta_i^j\}_{i \in \{i:\ \lambda_i > \lambda_0\}, j \in [-N:N]} \in \mathbb{R}^{(2N+1)|\{i:\ \lambda_i > \lambda_0\}|},$$
$$\{y_i\}_{i=1}^D \in \mathbb{R}^D, \quad s \in \mathbb{R}, \quad z \in \mathbb{R}^D.$$

This problem is a relaxation of the ℓ_1-norm relaxed problem (16). Besides, it holds that

$$\textit{optval}(\text{STRS}) \le \text{optval}(17) \le \rho^2 \textit{optval}(\text{STRS}) + (\rho^2 - \rho)\|t\|_2 + \frac{1}{4N^2} \sum_{i:\ \lambda_i > \lambda_0} (\lambda_i - \lambda_0)\theta_i^2,$$

where the constant $\rho > 0$ is defined in Theorem 5.

The convex integer program (17) seems appealing because it only requires solving $O((2N)^{|\{i:\ \lambda_i > \lambda_0\}|})$ number of finite-dimensional convex optimization problem to obtain its optimal solution. In practice, the choice of λ_0 influences the computational traceability of problem (17), and the choice of N influences the quality of the approximation. We follow the heuristic described in [17, Section 4.3.1] to select λ_0 and N. After solving the problem (17), one obtains the decision variable z that may not be feasible in $\mathcal{Z}$. Then, one can use the greedy rounding scheme to project z onto $\mathcal{Z}$ to obtain a primal feasible solution.

4. MMD Optimization with Different Kernels

In this section, we provide detailed algorithms for solving the MMD optimization problem (7) for various kernels considered in (4)-(6).

4.1. Linear Kernel Case

For linear kernel defined in (4), one can verify that $H_{i,j}$ defined in (2) can be written as a linear function in terms of z:

$$H_{i,j} = \sum_{s\in[D]} z^{(s)} \left[k_s(x_i^{(s)}, x_j^{(s)}) + k_s(y_i^{(s)}, y_j^{(s)}) - k_s(x_i^{(s)}, y_j^{(s)}) - k_s(y_i^{(s)}, x_j^{(s)}) \right] = z^{\mathrm{T}} h_{i,j},$$

where we denote the vector

$$h_{i,j} = (k_s(x_i^{(s)}, x_j^{(s)}) + k_s(y_i^{(s)}, y_j^{(s)}) - k_s(x_i^{(s)}, y_j^{(s)}) - k_s(y_i^{(s)}, x_j^{(s)}))_{s\in[D]}. \tag{18}$$

Since the empirical MMD estimator $\widehat{\mathrm{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$ is a linear combination of $\{H_{i,j}\}_{i,j}$ and the empirical variance estimator $\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n, \mathbf{y}^n; K_z)$ is a quadratic function in terms of $\{H_{i,j}\}_{i,j}$, it is clear that the MMD optimization problem (7) can be reformulated as a mixed-integer quadratic optimization problem, i.e., a STRS that has been studied in Section 3:

$$\mathsf{optval}(7) = \max_{z\in\mathcal{Z}} \left\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} t \right\},$$

where the data matrix $A \in \mathbb{R}^{D\times D}$ and $t \in \mathbb{R}^D$ have the following expressions:

$$A^{(s_1,s_2)} = \frac{4\lambda}{n^3} \sum_{i\in[n]} \left(\sum_{j\in[n]} h_{i,j}^{(s_1)} \right) \left(\sum_{j\in[n]} h_{i,j}^{(s_2)} \right) - \frac{4\lambda}{n^4} \left(\sum_{i,j\in[n]} h_{i,j}^{(s_1)} \right) \left(\sum_{i,j\in[n]} h_{i,j}^{(s_1)} \right), \quad \forall s_1, s_2 \in [D],$$
$$t = \sum_{i\in[n], j\in[n], i\neq j} h_{i,j}.$$

Therefore, one can query either the exact or approximation algorithm to solve problem (7) with strong optimization guarantees for this linear kernel case. In the following remark, we discuss under which conditions will linear kernel MMD may or may not achieve satisfactory performance on the variable selection task.

Remark 1 (Concerns about Linear Kernel). Under the linear kernel choice, it can be shown that

$$\mathrm{MMD}^2(\mu, \nu; K_z) = \sum_{s\in[D]} z^{(s)} \mathrm{MMD}^2(\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu; k_s)$$

where $\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu$ are the s-th marginal distributions of μ, ν, respectively. In other words, the selection coefficient z aims to find a direction to identify the difference between marginal distributions of μ and ν. However, under the case where marginal distributions of μ and ν are the same, the linear kernel MMD does not have enough power to find informative variables to distinguish those two distributions. ♣

4.2. Other Kernel Choices

For other kernel choices such as the quadratic kernel in (5) and Gaussian kernel in (6), the objective for MMD optimization is a nonlinear non-concave function with respect to z. This, together with the sparse constraint of the domain set $\mathcal{Z}$, makes this type of problem very challenging to solve. In this subsection, we provide a heuristic algorithm that incorporates simulated annealing (SA) and STRS that tries to find a feasible solution of (7) with high solution quality. Such a heuristic can also be naturally extended for generic kernel choices.

Here, we outline our SA and STRS-based heuristic. For notational simplicity, we denote the objective of (7) as $F(z)$ instead. Our proposed algorithm is an iterative method that generates a trajectory

Algorithm 3 Heuristic algorithm for solving (7) with generic kernel

Input: Max iterations $i_{\max}$, initial guess z_1, initial temperature Tem, cooling parameter α, and a set of regularization values $\mathcal{G}$
for $i = 1, \ldots, i_{\max} - 1$ **do**
  Randomly pick the regularization value τ_i from $\mathcal{G}$.
  Obtain $\tilde{z}_i$ by solving a STRS in (19).
  Compute residual level $\Delta_i = F(\tilde{z}_i) - F(z_i)$ and probability $p_i = e^{\Delta_i / Tem}$
  if $\mathrm{rand}(0,1) < p_i$ **then**
    $z_{i+1} = \tilde{z}_i$
  else
    $z_{i+1} = z_i$
  end if
  $Tem = \alpha \cdot Tem$
end for
Return $z_{i_{\max}}$

of feasible solutions $z_1, \ldots, z_{i_{\max}}$. At the iteration point z_i, we generate a candidate solution $\tilde{z}_i$ by optimizing a second-order approximation of the objective $F(z)$ with quadratic penalty regularization around z_i:

$$\tilde{z}_i = \arg\max_{z \in \mathcal{Z}} \left\{ F(z_i) + \nabla F(z_i)^{\mathrm{T}}(z - z_i) + \frac{1}{2}(z - z_i)^{\mathrm{T}} \nabla^2 F(z_i)(z - z_i) - \frac{\tau_i}{2}\|z - z_i\|_2^2 \right\}, \tag{19}$$

where τ_i denotes the quadratic regularization value. Such a problem is a special case of STRS, which can be solved by querying the exact or approximation algorithm described in Section 3. Let $\Delta(z_i, \tilde{z}_i) = F(\tilde{z}_i) - F(z_i)$ denote the residual value for moving from z_i to $\tilde{z}_i$. The central idea of SA is to always accept moves with positive residual values while not forbidding moves with negative residual values. Specifically, we assign a certain temperature Tem, and update z_{i+1} as $\tilde{z}_i$ according to the probability

$$p(z_i, \tilde{z}_i; \mathrm{Tem}) = \begin{cases} 1, & \text{if } \Delta(z_i, \tilde{z}_i) \geq 0 \\ e^{\Delta(z_i, \tilde{z}_i)/Tem}, & \text{if } \Delta(z_i, \tilde{z}_i) < 0. \end{cases}$$

If the candidate solution $\tilde{z}_i$ is not accepted, we update z_{i+1} as z_i. The temperature parameter Tem is a critical hyper-parameter in this algorithm. We assign an initial value of Tem and iteratively decrease it such that in the last iterations the moves with worse objective values are less and less likely to be accepted. See our detailed algorithm procedure in Algorithm 3.

Finally, we add remarks regarding the tractability and flexibility of quadratic and Gaussian kernels.

Remark 2 (Quadratic Kernel). For quadratic kernel defined in (5), it can be shown that

$$\mathrm{MMD}^2(\mu, \nu; K_z) = z^{\mathrm{T}} \mathcal{A}(\mu, \nu) z + z^{\mathrm{T}} \mathcal{T}(\mu, \nu),$$

where $\mathcal{A}(\mu, \nu)$ is a $\mathbb{R}^{D \times D}$-valued mapping such that

$$\begin{aligned}(\mathcal{A}(\mu, \nu))^{(s_1, s_2)} = {} & \mathbb{E}_{x, x' \sim \mu}[k_{s_1}(x^{(s_1)}, x'^{(s_1)}) k_{s_2}(x^{(s_2)}, x'^{(s_2)})] \\ & + \mathbb{E}_{y, y' \sim \nu}[k_{s_1}(y^{(s_1)}, y'^{(s_1)}) k_{s_2}(y^{(s_2)}, y'^{(s_2)})] - 2\mathbb{E}_{x \sim \mu, y \sim \nu}[k_{s_1}(x^{(s_1)}, y^{(s_1)}) k_{s_2}(x^{(s_2)}, y^{(s_2)})],\end{aligned}$$

and $\mathcal{T}(\mu, \nu)$ is a $\mathbb{R}^D$-valued mapping such that

$$(\mathcal{T}(\mu, \nu))^{(s)} = 2c\mathrm{MMD}^2(\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu; k_s).$$

Given two multivariate distributions, the quadratic MMD aims to find a direction z to distinguish the difference in each coordinate together with the correlation between two coordinates the most.

Compared with the linear MMD that only identifies the difference in each coordinate, the quadratic MMD is a more flexible choice. However, it can be shown that the objective in (7) with the quadratic kernel is a 4-th order non-concave monomial with respect to z, which is computationally intractable to optimize. In practical experiments, we use the heuristic algorithm in Algorithm 3 to obtain a reasonably high-quality solution. ♣

REMARK 3 (GAUSSIAN KERNEL). One can also re-write the population testing statistic for the Gaussian kernel defined in (6). For notational simplicity, let $K(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right), \forall x, y \in \mathbb{R}^d$ be a standard Gaussian kernel with low-dimensional data, and define $z_{\#}\nu$ as a d-dimensional distribution such that

$$z_{\#}\nu = \left(z^{(s)}x^{(s)}\right)_{s\in\mathrm{supp}(z)}, \quad \text{where } x \sim \nu.$$

With these notations, it can be shown that

$$\mathrm{MMD}^2(\mu,\nu;K_z) = \mathrm{MMD}^2(z_{\#}\mu, z_{\#}\nu; K).$$

Because the kernel K satisfies the universal property [44], our proposed Gaussian kernel distinguishes the difference between μ and ν as long as there exists a d-size sub-group of coordinates of μ and ν that cause the difference. Compared with linear and quadratic kernels, the Gaussian kernel is a more flexible choice. Unfortunately, the computation burden of the Gaussian kernel is heavier than the other two simple kernels because the objective in (7) can be viewed as a non-concave ∞-degree monomial with respect to z, whereas the second-order approximation scheme in (19) may not provide reliable performance for optimization. ♣

5. Testing Power Analysis

In this section, we provide statistical performance guarantees for the variance-regularized MMD statistics in (7). We first make the following assumptions regarding the kernel choice $K_z(\cdot,\cdot)$, variance regularization value λ, and data distributions μ, ν.

ASSUMPTION 1. *The kernel $K_z(\cdot,\cdot)$ is uniformly bounded and satisfies the Lipshitz continuous condition, i.e., for any $z, z' \in \mathcal{Z}, x, y \in \Omega$, it holds that $|K_z(x,y)| \le M$ and $|K_z(x,y) - K_{z'}(x,y)| \le L\|z - z'\|_2$.*

ASSUMPTION 2. *Under the alternative hypothesis $\mathcal{H}_1: \mu \neq \nu$, there exists $\lambda \ge 0$ such that for some $\bar{z} \in \mathcal{Z}$, it holds that $\mathrm{MMD}^2(\mu,\nu;K_{\bar{z}}) > 0$ and*

$$\Delta_{\bar{z}} \triangleq \mathrm{MMD}^2(\mu,\nu;K_{\bar{z}}) - \lambda\big[\max_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) - \min_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\big] > 0. \tag{20}$$

Here $\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)$ denotes the population version of the empirical variance estimator defined in (8).

Assumption 1 is a standard assumption used in the statistical analysis of kernel-based testing in literature. Besides, it is worth noting that Assumption 2 is not too restrictive: by properly specifying the kernel function $K_{\bar{z}}(\cdot,\cdot)$, one can make the population testing statistic $\mathrm{MMD}^2(\mu,\nu;K_{\bar{z}})$ strictly positive. For example, when some marginal distributions of μ, ν are different, according to Remark 1, the linear kernel satisfies this technical condition. Besides, one can impose a uniform bounded condition of the kernel $\{K_z(\cdot,\cdot)\}_{z\in\mathcal{Z}}$ to argue that $\max_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) - \min_{z\in\mathcal{Z}} \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)$ is also bounded. Hence, we can take a sufficiently small value of λ to make the condition (20) in Assumption 2 satisfied. In the following, we demonstrate that, under mild conditions, our proposed kernels in (4)-(6) indeed satisfy Assumptions 1 and 2.

PROPOSITION 2 **(Sufficient Condition of Assumptions 1 and 2).** (I) *(Linear Kernel) For the kernel in (4), Assumption 1 is guaranteed to hold with $M = \sqrt{d}$ and $L = \sqrt{2d}$. As long as there exists $s^* \in [D]$ such that $\mathrm{Proj}_{s^*\#}\mu \neq \mathrm{Proj}_{s^*\#}\nu$, Assumption 2 is guaranteed to hold with*

$$\lambda \in \left[0, \frac{\max_{z\in\mathcal{Z}} \sum_{s\in[D]} z^{(s)} \mathrm{MMD}^2(\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu; k_s)}{16d}\right).$$

(II) *(Quadratic Kernel) For the kernel in (5), Assumption 1 is guaranteed to hold with* $M = 2c^2 + 2d$ *and* $L = 4d + 2c\sqrt{2d}$*. As long as there exists* $s^* \in [D]$ *such that* $\mathrm{Proj}_{s^*\#}\mu \neq \mathrm{Proj}_{s^*\#}\nu$ *or* $(\mathcal{A}(\mu,\nu))^{(s^*,s^*)} > 0$ *holds, Assumption 2 is guaranteed to hold with*

$$\lambda \in \left[0, \frac{\max_{z\in\mathcal{Z}} z^{\mathrm{T}}\mathcal{A}(\mu,\nu)z + z^{\mathrm{T}}\mathcal{T}(\mu,\nu)}{16(2d+2c^2)^2}\right).$$

(III) *(Gaussian Kernel) For the kernel in (6), if additionally assuming that* $\Omega \subseteq \{x \in \mathbb{R}^D : \|x\|_\infty \leq R\}$*, Assumption 1 is guaranteed to hold with* $M = 1$ *and* $L = \frac{2R}{\sigma\sqrt{e}}$*. As long as there exists* $S \subseteq [D]$ *with* $|S| \leq d$ *such that* $\mathrm{Proj}_{S\#}\mu \neq \mathrm{Proj}_{S\#}\nu$*, Assumption 2 is guaranteed to hold with*

$$\lambda \in \left[0, \frac{\max_{z\in\mathcal{Z}} \mathrm{MMD}^2(\mu,\nu;K_z)}{16}\right).$$

Next, we derive concentration properties to show that, the empirical estimators $S^2(\mathbf{x}^n,\mathbf{y}^n;K_z)$ and $\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)$ uniformly converge to their population version as the sample size n increases. Such a property is useful for showing the testing consistency and the rate of testing power of our MMD framework.

THEOREM 6 **(Non-asymptotic Concentration Properties)**. *Under Assumption 1, with probability at least* $1-\delta$*, (i) the bias approximation error can be bounded as*

$$\sup_{z\in\mathcal{Z}}\left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| \leq \epsilon^1_{n,\delta} \triangleq \frac{8}{\sqrt{n}}\left[M\sqrt{2\log\binom{D}{d}\frac{2}{\delta} + 2d\log(4\sqrt{n})} + L\right],$$

(ii) and the variance approximation error can be bounded as

$$\sup_{z\in\mathcal{Z}}\left|\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \leq \epsilon^2_{n,\delta} \triangleq \frac{64}{\sqrt{n}}\left[7\sqrt{2\log\binom{D}{d}\frac{2}{\delta} + 2d\log(4\sqrt{n})} + \frac{18M^2}{\sqrt{n}} + 8LM\right],$$

where $\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) \triangleq \mathbb{E}_{\mathbf{x}^n\sim\mu,\mathbf{y}^n\sim\nu}[\widehat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z)]$.

Proof of the theorem above follows similar covering number arguments in [40, Theorem 6]. The main difference is that when applying union bound on the set $\mathcal{Z}$, the corresponding error bound is sharper because the covering number of sparse-constrained set $\mathcal{Z}$ is much smaller. Finally, we are ready to prove the main theorem of this section. For fixed $\delta \in (0,1)$ and sample size n, we define the following error parameter for notational simplicity:

$$\epsilon_{n,\delta} = \epsilon^1_{n,\delta} + \lambda\epsilon^2_{n,\delta}.$$

As a simple corollary from Theorem 6, the objective in (7) approximates its population version with error $\epsilon_{n,\delta/2}$ with probability at least $1-\delta$.

THEOREM 7 **(Asymptotic Distribution of Testing Statistic)**. *Under Assumptions 1 and 2, let* $\hat{z}_{\mathrm{Tr}}$ *be the obtained sparse coefficient by solving (7) from training dataset* $(\mathbf{x}^{\mathrm{Tr}},\mathbf{y}^{\mathrm{Tr}})$ *with* $|\mathbf{x}^{\mathrm{Tr}}| = |\mathbf{y}^{\mathrm{Tr}}| = n_{\mathrm{Tr}}$*, and* $T_{n_{\mathrm{Te}}}$ *be the testing statistic evaluated on testing dataset* $(\mathbf{x}^{\mathrm{Te}},\mathbf{y}^{\mathrm{Te}})$ *with* $|\mathbf{x}^{\mathrm{Te}}| = |\mathbf{y}^{\mathrm{Te}}| = n_{\mathrm{Te}}$*. Then, it holds that*

(I) *Under alternative hypothesis* $\mathcal{H}_1 : \mu \neq \nu$*,* $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] \geq \Delta_{\hat{z}} - 2\epsilon_{n_{\mathrm{Tr}},\delta/4}$ *with probability at least* $1-\delta$*, where the expectation is taken with respect to the randomness from testing dataset, and the probability error is from the randomness from training dataset. In other words, when training sample size* n_{Tr} *is sufficiently large so that the error* $\epsilon_{n_{\mathrm{Tr}}}$ *is sufficiently small, we have that* $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] > 0$.

(II) *Under null hypothesis* $\mathcal{H}_0: \mu = \nu$, *it holds that* $n_{\mathrm{Te}} T_n \to \sum_i \sigma_i (Z_i^2 - 2)$, *with* σ_i *denoting the eigenvalues of the* μ*-covariance operator of the centered kernel; under alternative hypothesis* $\mathcal{H}_1: \mu \neq \nu$, *it holds that* $\sqrt{n_{\mathrm{Te}}}(T_n - \mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1]) \to \mathcal{N}(0, \sigma^2_{\mathcal{H}_1}(\mu, \nu; K_{\hat{z}_{\mathrm{Tr}}}))$.

COROLLARY 2 **(Consistency)**. *Under the setting of Theorem 7, suppose that* $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] > 0$. *Let* $\alpha \in (0,1)$ *denote the level of two-sample test and take* τ *as the* $(1-\alpha)$*-quantile of the limiting distribution* $\sum_i \sigma_i (Z_i^2 - 2)$ *defined in Theorem 7(II), and let the threshold of the test be* $t_{\mathrm{thres}} := \frac{\tau}{n_{\mathrm{Te}}}$. *As a consequence,*

$$\mathbb{P}\big(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}} \mid \mathcal{H}_0\big) \to \alpha, \quad \mathbb{P}\big(T_{n_{\mathrm{Te}}} \leq t_{\mathrm{thres}} \mid \mathcal{H}_1\big) \to 0.$$

The consistency of the testing framework requires the assumption that $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] > 0$, i.e., $\epsilon_{n_{\mathrm{Tr}},\delta/4}$ should be a sufficiently small number. It is worth mentioning that $\epsilon_{n_{\mathrm{Tr}},\delta/4} = O(1/\sqrt{n_{\mathrm{Tr}}})$, where the precise hidden constant relies on factors $\mathrm{poly}(D), \mathrm{poly}(d), \mathrm{polylog}(1/\delta)$, and $\mathrm{polylog}(n_{\mathrm{Tr}})$, indicating the statistical guarantees of our proposed variable selection framework do not suffer from the curse of dimensionality.

THEOREM 8 **(Testing Power)**. *Under the setting of Theorem 7, suppose that* $\mathbb{E}[|T_1|^3 \mid \mathcal{H}_1] < \infty$ *and* $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] > 0$. *Let* $\alpha \in (0,1)$ *denote the level of two-sample test and take* τ *as the* $(1-\alpha)$*-quantile of the limiting distribution* $\sum_i \sigma_i (Z_i^2 - 2)$ *defined in Theorem 7(II), and let the threshold of the test be* $t_{\mathrm{thres}} := \frac{\tau}{n_{\mathrm{Te}}}$. *When the testing sample size* n_{Te} *is sufficiently large so that*

$$\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1] \geq t_{\mathrm{thres}} + \frac{\Phi^{-1}(1 - n_{\mathrm{Te}}^{-1/2})}{\sqrt{n_{\mathrm{Te}}}} = \frac{\tau}{n_{\mathrm{Te}}} + \sqrt{\frac{\ln \frac{n_{\mathrm{Te}}}{2\pi} - \ln\ln \frac{n_{\mathrm{Te}}}{2\pi}}{n_{\mathrm{Te}}}}(1 + o(1)),$$

it holds that

$$\mathbb{P}\big(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}} \mid \mathcal{H}_0\big) \leq \alpha + O(n_{\mathrm{Te}}^{-1/2}), \quad \mathbb{P}\big(T_{n_{\mathrm{Te}}} \leq t_{\mathrm{thres}} \mid \mathcal{H}_1\big) \leq O(n_{\mathrm{Te}}^{-1/2}),$$

where $O(\cdot)$ *hides constant related to parameters* $\mathbb{E}[|T_1|^3 \mid \mathcal{H}_1]$ *and* $\sigma^2_{\mathcal{H}_1}$.

Theorem 8 indicates that under alternative hypothesis $\mu \neq \nu$, as long as the testing sample size n_{Te} is sufficiently large such that $\mathbb{E}[T_{n_{\mathrm{Te}}} \mid \mathcal{H}_1]$ dominates $n_{\mathrm{Te}}^{-1/2}$ multiplied by a near-constant factor $\tilde{O}(1)$, the testing power approaches 1 with error rate $O(n_{\mathrm{Te}}^{-1/2})$. Besides, Theorem 7 ensures that $\mathbb{E}T_{n_{\mathrm{Te}}} = \Theta(1)$ with probability at least $1 - \delta$ as long as $\frac{n_{\mathrm{Tr}}}{\mathrm{polylog}(n_{\mathrm{Tr}})} = \Omega(1/\Delta_{\hat{z}}^2)$, where the precise hidden constant relies on factors $\mathrm{poly}(D), \mathrm{poly}(d)$, and $\mathrm{polylog}(1/\delta)$. Combining those two theorems, we imply that our proposed MMD test achieves satisfactory performance as long as both the training and testing sample sizes are moderately large.

6. Numerical Simulation

We first consider synthesized data sets to examine the performance of our proposed variable selection framework. We consider the following four cases:

(I) (Gaussian Mean Shift): Data distribution $\mu = \mathcal{N}(0, \Sigma)$ with the covariance matrix $\Sigma^{(s_1,s_2)} = \rho^{|s_1 - s_2|}$ for some correlation level $\rho \in (0,1)$. Data distribution $\nu = \mathcal{N}(\mu, \Sigma)$ with the mean vector $\mu^{(s)} = \tau/s, \forall s \in [d_{\mathrm{true}}]$ for some scalar $\tau > 0$ and otherwise $\mu^{(s)} = 0$.

(II) (Gaussian Covariance Shift): Data distribution $\mu = \mathcal{N}(0, \Sigma)$ with Σ specified the same as in Part (I), and $\nu = \mathcal{N}(0, \tilde{\Sigma})$, with $\tilde{\Sigma}^{(s_1,s_2)} = \tau \Sigma^{(s_1,s_2)}, \forall s_1, s_2 \in [d_{\mathrm{true}}]$ for some scalar $\tau > 1$ and otherwise $\tilde{\Sigma}^{(s_1,s_2)} = \Sigma^{(s_1,s_2)}$.

(III) (Gaussian versus Laplacian): Data distribution $\mu = \mathcal{N}(0, I_D)$. The first d_{true} coordinates of ν are independent Laplace distributions with zero mean and standard deviation 0.8. The remaining coordinates of ν_Y are independent Gaussian distributions $\mathcal{N}(0,1)$.

(IV) (Gaussian versus Gaussian Mixture): Data distribution $\mu = \mathcal{N}(0, I_D)$. The first d coordinates of ν are Gaussian mixture distribution $\frac{1}{2}\mathcal{N}(-\mu, I_{d_{\text{true}}}) + \frac{1}{2}\mathcal{N}(\mu, I_{d_{\text{true}}})$ while the remaining coordinates are independent Gaussian distributions $\mathcal{N}(0,1)$. Here the mean vector $\mu^{(s)} = \tau/s, \forall s \in [d_{\text{true}}]$ for some scalar $\tau > 0$.

Throughout the numerical experiment, we take hyper-parameters in Case (I) as $\tau = 1, \rho = 0.5$, in Case (II) as $\tau = 2, \rho = 0.5$, and in Case (IV) as $\tau = 2$. We quantify the performance in terms of hypothesis testing metrics rather than the prediction accuracy metrics used in the literature. Besides, we also measure the quality of variable selection using *false-discovery proportion* (FDP) and the *non-discovery proportion* (NDP) defined in [3]:

$$\text{FDP}(I) = \frac{|I \setminus I^*|}{|I|}, \quad \text{NDP}(I) = \frac{|I^* \setminus I|}{|I^*|}, \tag{21}$$

where I^* denotes the ground truth feature set and I denotes the set obtained by variable selection algorithms. The smaller the FDP or NDP is, the better performance the obtained feature set has.

For simplicity of implementation, we chose the bandwidth hyper-parameter τ_s^2 for the kernel $k_s(x, y)$ using the median heuristic, i.e., we specify it as the median among all pairwise distances for data points in the s-th coordinate. Similarly, we take bandwidth σ^2 of Gaussian kernel as the median among all pairwise distances for data points (over all coordinates), and bandwidth of quadratic kernel as $c = \sqrt{\sigma^2}$. Users are also recommended to tune those hyper-parameters based on the cross-validation technique, which tends to return near-optimal hyper-parameter choices for large sample sizes.

6.1. Numerical Performance for Solving (STRS)

We first examine the numerical performance of various approximation algorithms for solving (STRS), by taking the MMD optimization with linear kernel (see the reformulation in Section 4.1) as a numerical example. For each of the four synthetic datasets, we try various choices of parameters (N, D, d) from the set

$$\{(1\text{e}3, 20, 5), (1\text{e}3, 40, 10), (1\text{e}3, 60, 6), (1\text{e}3, 80, 8), (1\text{e}3, 100, 10)\}.$$

We also specify different hyper-parameters $\lambda \in \{0.8, 0.7, 0.6, 0.5\}$ when using these four different datasets, respectively. Since those approximation algorithms may return a solution that is infeasible to the constraint $\mathcal{Z}$, we estimate the corresponding feasible solution by performing the normalized sparse truncation (see Definition 2). Figure 1 reports the objective value obtained from the feasible solution based on those approximation algorithms, where the error bars are generated using 100 independent trials. The larger the objective value is, the better performance the designed algorithm has. From the plot, we can check that semidefinite relaxation and convex integer programming algorithms have nearly optimal performance compared with the ground truth, whereas the performance truncation algorithm is slightly worse compared with those approaches. Table 1 reports the corresponding computation time of those approximation algorithms, from which we identify that the truncation algorithm has the fastest computational speed while SDP relaxation has the slowest speed. Since the convex integer programming algorithm has satisfactory performance with relatively fast computational speed, we recommend using this approximation algorithm when solving (STRS).

6.2. Impact of Sample Size and Data Dimension

In this subsection, we compare the performance of variable selection based on the following approaches: (I) Linear kernel MMD; (II) Quadratic kernel MMD; (III) Gaussian kernel MMD; (IV) Sparse Logistic Regression: a framework that trains the projection vector with ℓ_0-norm constraint to minimize the logistic loss [8]; and (V) Projected Wasserstein: variable selection framework using projected Wasserstein distance [47]. For baselines (I)-(III), we also compare the performance of standard

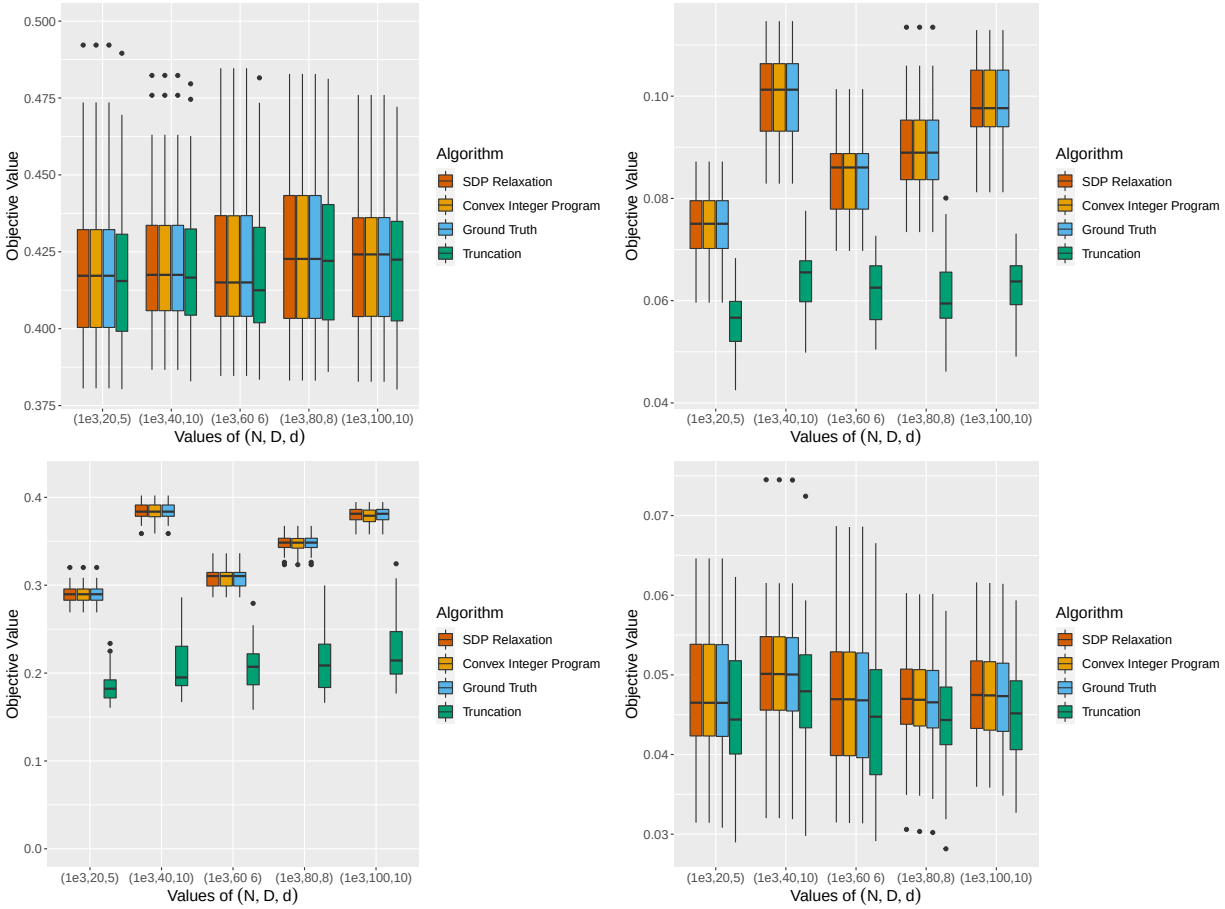

Figure 1 Box plots on the performance of various approximation algorithms for solving (STRS). The x-axis corresponds to various choices of (N, D, d), and y-axis corresponds to the estimated objective value of (STRS). Plots from top to bottom correspond to four types of synthetic datasets.

MMD testing without the variable selection technique. We quantify the performance using the testing power metric with controlled type-I error $\alpha_{\text{level}} = 0.05$, and take the training/testing sample sizes as $n_{\text{Tr}} = n_{\text{Te}} = n$.

Figure 2 reports a numerical study on the impact of sample size n with data dimension $D = 100$, number of different variables $d_{\text{true}} = 20$ and sparsity level $d = 20$. The error bars are generated using 20 independent trials. From these plots, we find the sparse logistic regression does not have competitive performance in general. The explanation is that a linear classifier is not flexible enough to distinguish the distributions from two groups. Following the similar argument from Example 1, one can check the testing statistic of this baseline always equals to zero as long as the mean vectors of two distributions are the same, which explains why this baseline has nearly zero power for the synthetic dataset of case (II)-(IV). The testing power for the other two-sample testing methods increases with respect to the sample size. We can see the variable selection technique improves the performance of the standard MMD framework. For the first three synthetic datasets, the linear or quadratic MMD testing with variable selection achieves superior performance than other baselines, while for the last example, the projected Wasserstein distance has the best performance. One possible explanation is that the MMD testing framework may not be good at detecting distribution changes for Gaussian mixture distributions.

Next, we examine the impact of the data dimension D with fixed $n = 50, d_{\text{true}} = 20, d = 20$ in Figure 3. We omit to report the performance of the sparse logistic regression baseline because it does

Table 1 Averaged computational time of various approximation algorithms for solving (STRS).

Data Type	Parameters			Averaged Computational Time(s) of Approximation Algorithms		
	n	D	d	Truncation Algorithm	SDP Relaxation	Convex Integer Programming
Gaussian Mean Shift	1e3	20	5	2.13e-3	1.18	1.25e-1
	1e3	40	10	6.08e-3	2.55	2.29e-1
	1e3	60	6	1.30e-2	4.80	4.47e-1
	1e3	80	8	3.08e-2	6.19	6.87e-1
	1e3	100	10	6.87e-2	9.47	8.77e-1
Gaussian Covariance Shift	1e3	20	5	2.33e-3	1.18	1.24e-1
	1e3	40	10	5.86e-3	2.57	3.11e-1
	1e3	60	6	1.32e-2	4.80	4.02e-1
	1e3	80	8	3.07e-2	6.46	9.38e-1
	1e3	100	10	6.76e-2	9.73	1.23
Gaussian versus Laplacian	1e3	20	5	2.28e-3	1.29	1.69e-1
	1e3	40	10	6.39e-3	2.85	5.65e-1
	1e3	60	6	1.44e-2	5.20	5.14e-1
	1e3	80	8	3.31e-2	6.79	1.15
	1e3	100	10	6.95e-2	1.02e+1	2.10
Gaussian versus Gaussian Mixture	1e3	20	5	2.17e-3	1.16	1.17e-1
	1e3	40	10	6.38e-3	2.57	2.14e-1
	1e3	60	6	1.42e-2	4.72	3.94e-1
	1e3	80	8	3.31e-2	6.07	5.91e-1
	1e3	100	10	7.25e-2	9.44	8.67e-1

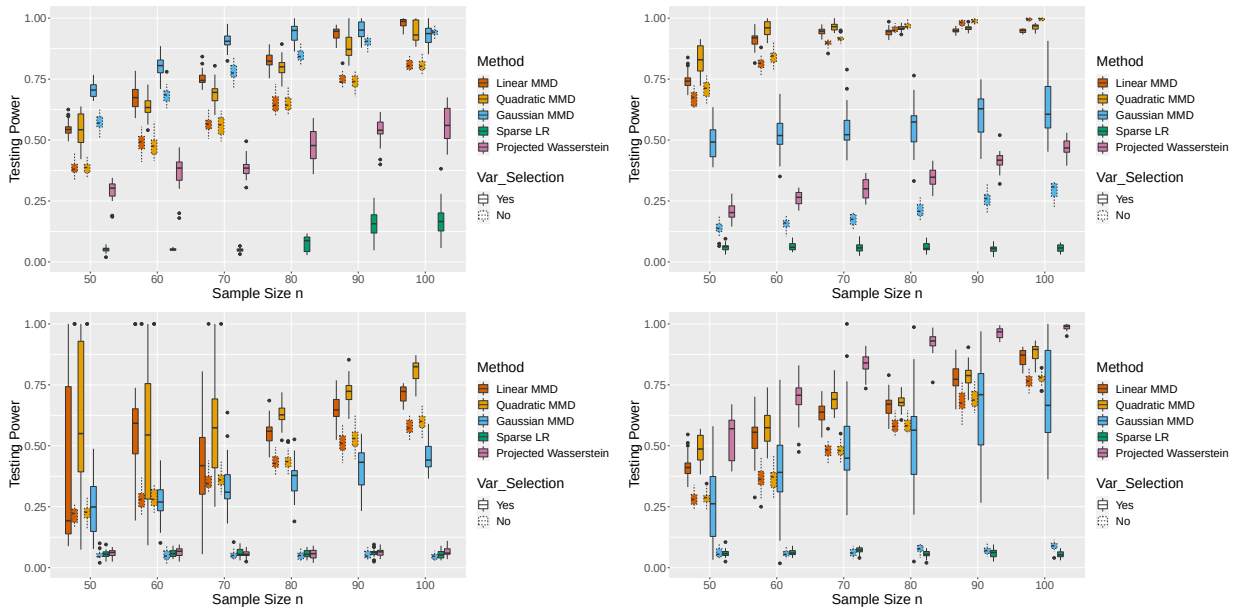

Figure 2 Testing power of various two-sample tests with different choices of sample size n. Here we fix parameters $D = 100,\ d_{\text{true}} = 20,\ d = 20$ and control the type-I error $\alpha_{\text{level}} = 0.05$. Plots from top to bottom correspond to four different types of synthetic datasets.

not achieve satisfactory testing performance as studied before. From those plots, we find that as the data dimension increases, all methods tend to have decreasing testing power. However, the decaying rate of MMD testing with the variable selection procedure seems to be slower than that of standard MMD testing. For the synthetic dataset of case (I), the Gaussian kernel has the best performance, while for case (II)-(III), the linear or quadratic kernel has the best performance. A possible explanation is that one can optimize the linear kernel with strong performance guarantees, whereas we only use quadratic approximation heuristics to optimize other types of kernel functions. Since the quadratic approximation of the objective for the quadratic kernel seems to be tight, it is intuitive to see the performance of the quadratic kernel is also consistently good.

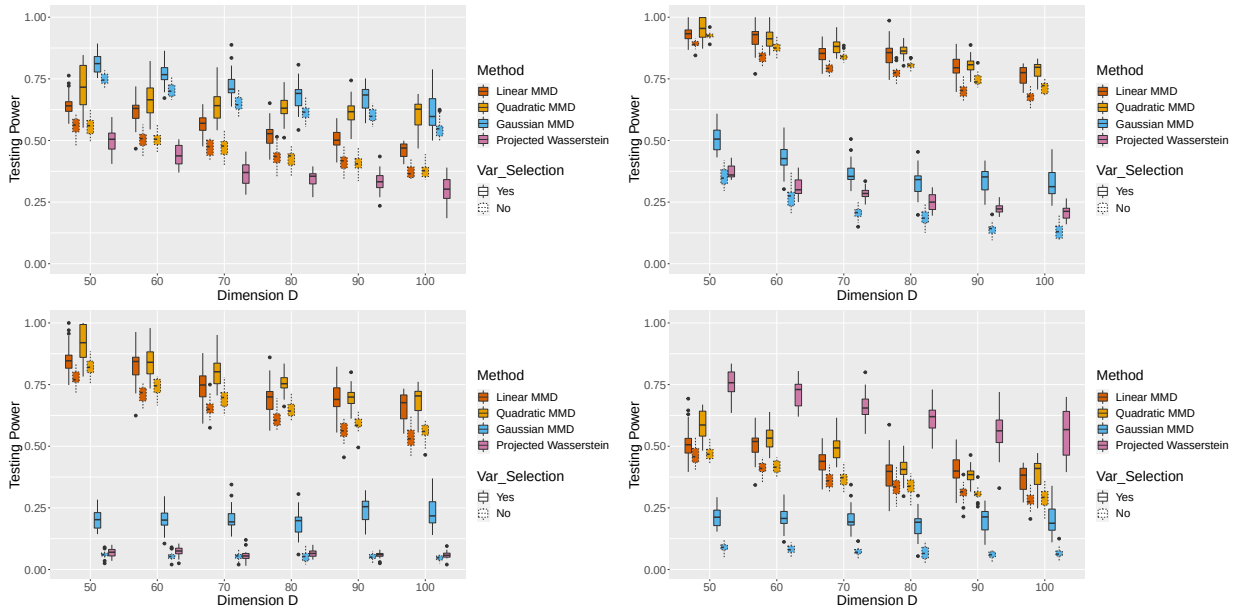

Figure 3 Testing power of various two-sample tests with different choices of data dimension D. Here we fix parameters $n = 50,\ d_{\text{true}} = 20,\ d = 20$ and control the type-I error $\alpha_{\text{level}} = 0.05$. Plots from top to bottom correspond to four different types of synthetic datasets.

6.3. Results on Support Recovery

In this subsection, we show the performance of support recovery for various variable selection approaches. We specify parameters $n = 100, D = 100, d_{\text{true}} = 20$ and vary the sparsity level $d = 1, \ldots, 40$. We quantify the performance using the FDP and NDP metrics defined in (21). Besides, we are interested in whether those methods perform consistently well in both metrics, so we further quantify the performance as the sum of those two metrics. Figure 4 presents these three metrics based on various variable selection approaches across different choices of sparsity level d for a single independent trial. From these plots, we can see that our proposed variable selection framework with linear or quadratic kernel achieves the best performance for all four cases, as indicated by the lowest values of FDP and NDP metrics. This observation is also consistent with the testing power performance examined in the previous subsection.

7. Numerical Study on Real-World Datasets

In this section, we present additional numerical study with real-world datasets. Specifically, we demonstrate a visualization of variable selection based on the MNIST handwritten digits image dataset in

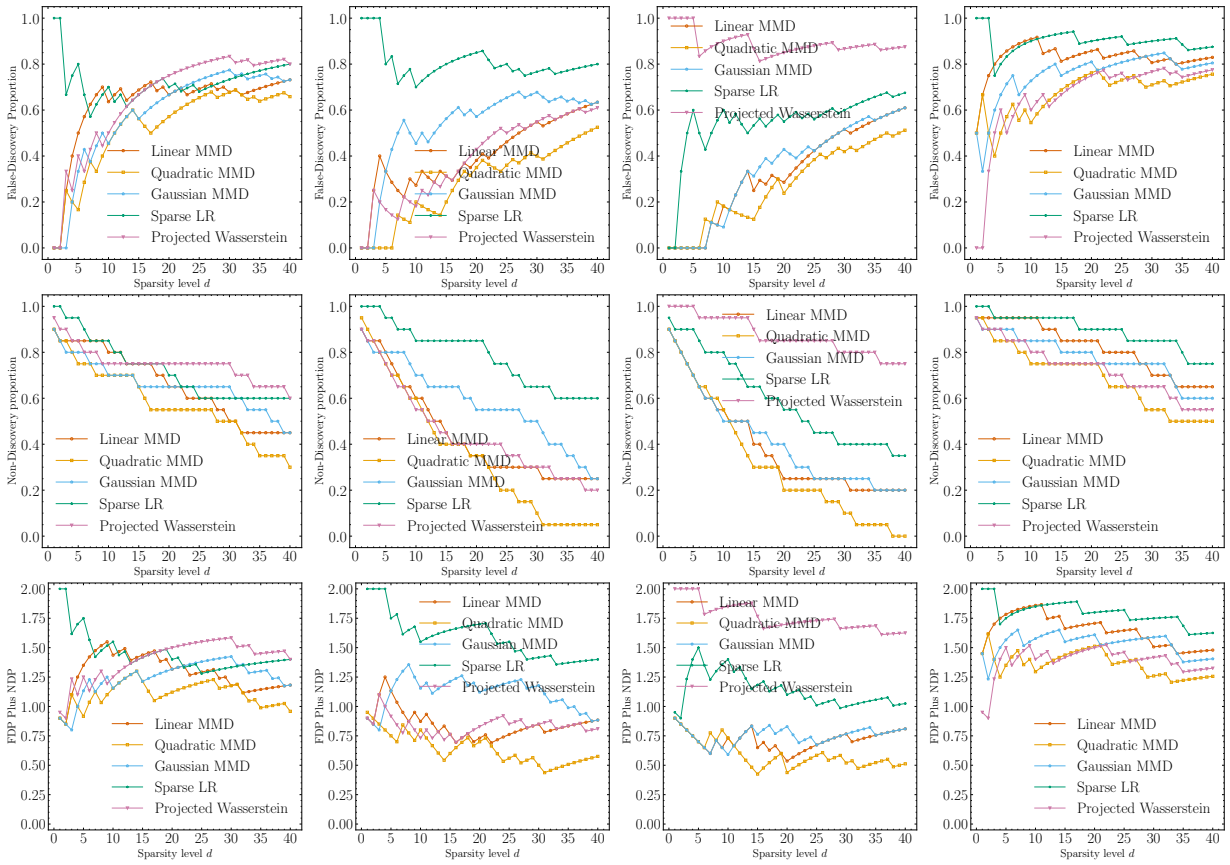

Figure 4 FDP and NDP metrics obtained by various approaches for different choices of sparity level d using synthetic datasets.

Section 7.1. Next, we show that the variable selection approach can help with identifying key variables for disease diagnosis in Section 7.2.

7.1. Visualization on MNIST Image Datasets

In this part, we demonstrate a visualization of our variable selection framework by taking the classification of MNIST image datasets, consisting of 28×28 gray-scale handwritten images for digits from 0 to 9, as toy examples. We take the training sample size $n_{\mathrm{Tr}} = 20$ and testing sample size $n_{\mathrm{Te}} = 5$. We pre-process the MNIST images by performing a 2d convolutional operator using the kernel of size 9×9. The pre-processed samples have dimension $D = 169$, and we take the number of selected variables $d = 20$. We construct four types of data distributions (μ, ν) for two-sample testing: μ and ν are distributions of images corresponding to digits 0 and 6, 8 and 9, 3 and 8, or 7 and 9, respectively. We show the visualization results in Figure 5. Specifically, different rows correspond to different data distributions for two-sample testing. Plots in the left two columns visualize the selected pixels (highlighted with red square markers) on two different image samples based on our linear kernel variable selection framework, from which we can see that our proposed method identifies the difference between two digits correctly. Plots in the 3rd column report the MMD statistic compared with the empirical distribution under H_0 via test-only bootstrap, where the green circle markers correspond to the bootstrap threshold for rejecting H_0 and red star markers correspond to the testing statistics. From these plots, we find our proposed framework has satisfactory testing power even with small training and testing sample sizes. Plots on the 4th column report the visualization of the distribution of the MNIST dataset after variable selection embedded in 2D generated by tSNE [58], which

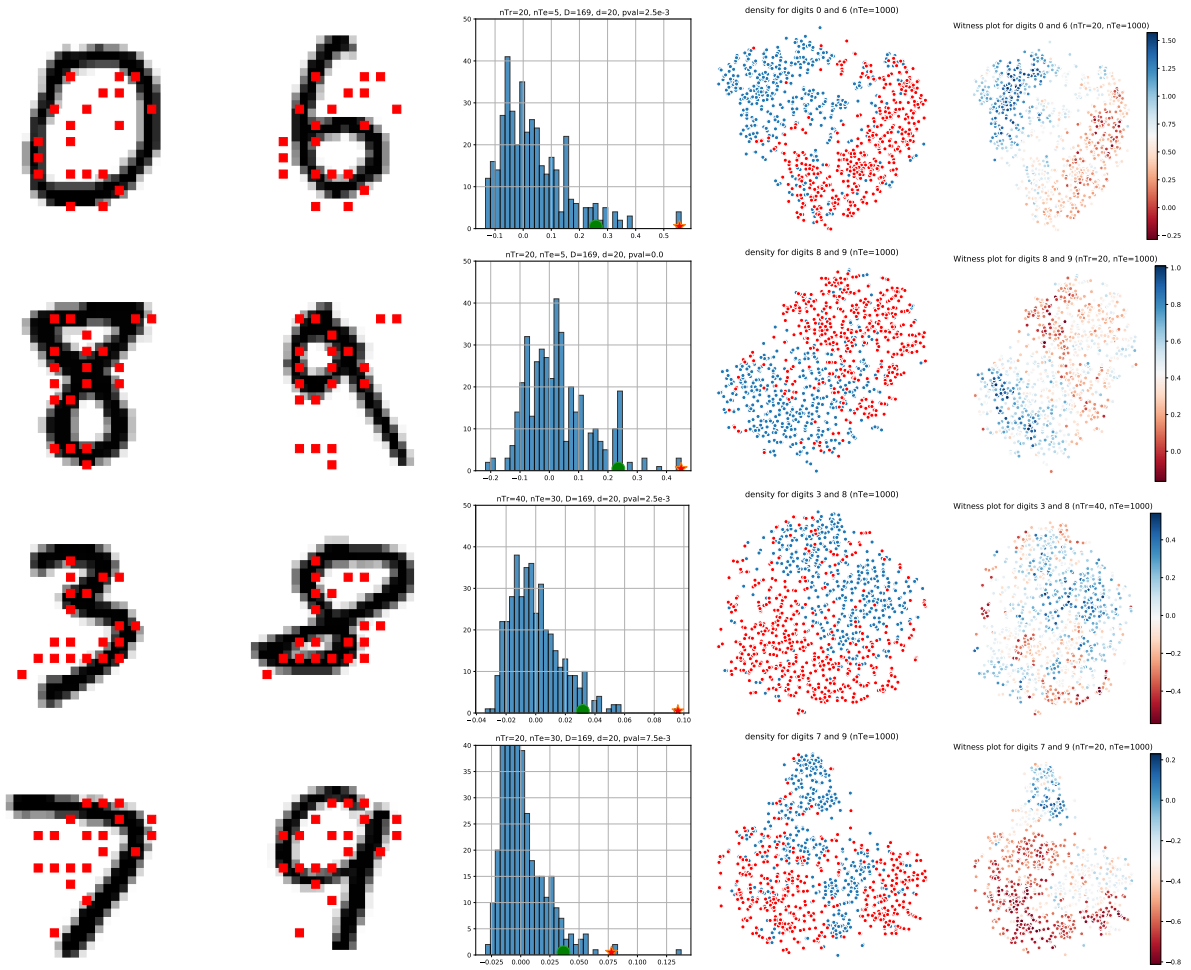

Figure 5 Different rows correspond to two-sample testing with different MNIST digits. The first two column plots visualize the selected pixels based on the variable selection framework. The 3rd column plots visualize the MMD statistic together with the empirical distribution under H_0 that is estimated via bootstrapping. The 4th column plots visualize the distribution of MNIST digits after variable selection embedded in 2D. The 5th column plots visualize the estimated witness function.

is estimated based on 1000 testing samples. In comparison, we plot the estimated witness function as a color field over those samples in the right-hand-side figures. From those plots, we can see that the estimated witness function identifies the region of the distribution change for all of these four two-sample testing tasks.

7.2. Healthcare Datasets

Finally, we study the performance of variable selection on a healthcare dataset [62] that records information for healthy people and Sepsis patients. This dataset consists of $D = 39$ features from $m = 20771$ healthy people and $n = 2891$ Sepsis patients. We take training samples with sample sizes $m_{\text{Tr}} = 20000, n_{\text{Tr}} = 2000$ and specify the remaining as validation samples. We quantify the performance of variable selection as the testing power on testing samples with sample size $m_{\text{Te}} = n_{\text{Te}} = 100$, which are selected randomly from the validation sample sets. We repeat the testing procedure for 2000 independent trials and report the average testing power in Table 2.

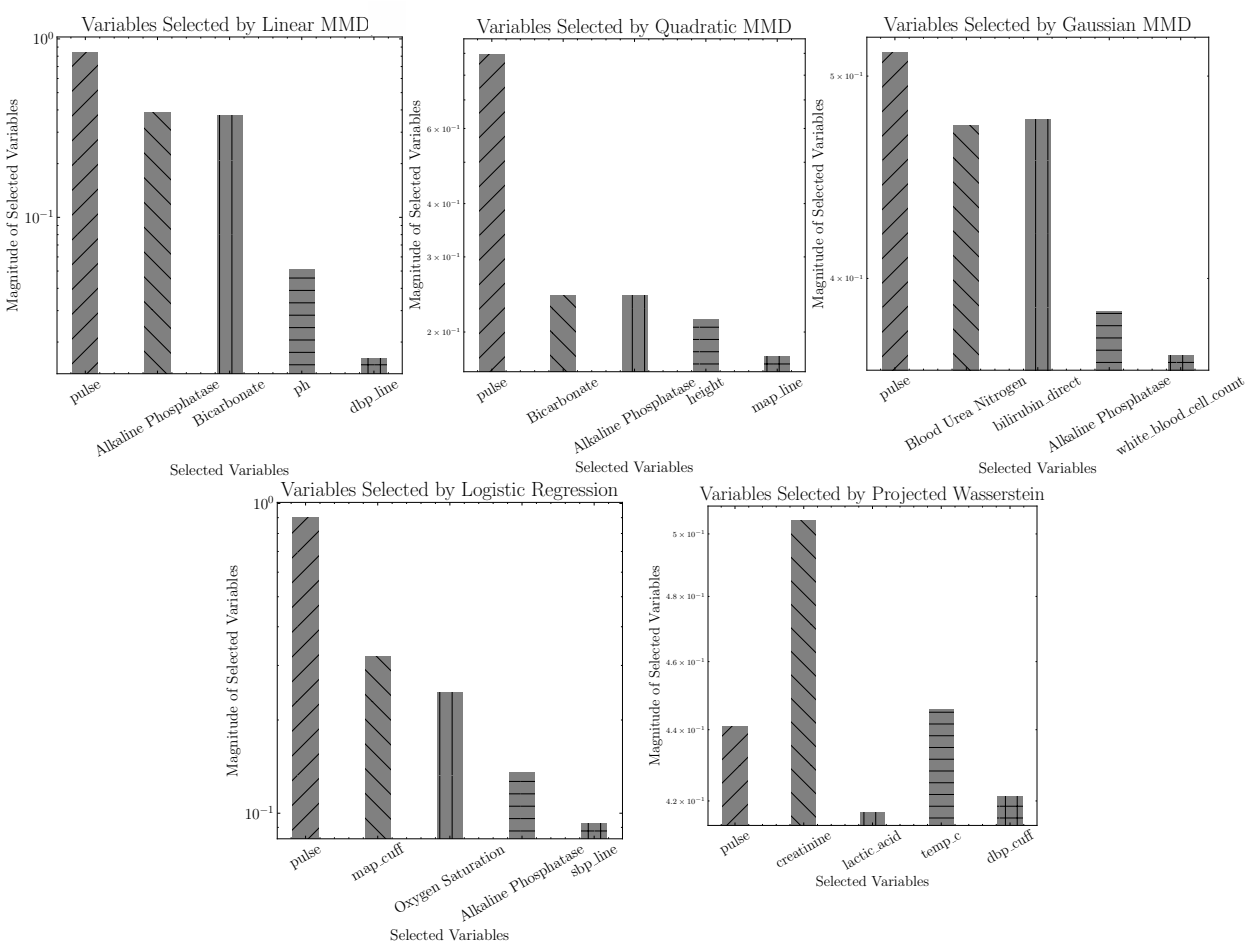

Figure 6 Top 5 variables selected by various approaches in the healthcare dataset.

We report the top 5 features selected by various approaches based on the training samples in Figure 6. From the Table, we can see that methods Quadratic MMD and Linear MMD perform the best, and the intersection of those selected features are pulse, Bicarbonate, Alkaline Phosphatase.

Table 2 Averaged testing power for the sepsis prediction.

Linear MMD	Quadratic MMD	Gaussian MMD	Logistic Regression	Projected Wasserstein
0.835	0.915	0.784	0.771	0.749

8. Conclusion

We studied variable selection for the kernel-based two-sample testing problem, which can be formulated as mixed-integer programming problems. We developed exact and approximate algorithms with performance guarantees to solve those formulations. Theoretical properties for the proposed frameworks are provided. Finally, we validated the power of this approach in synthetic and real datasets. In the meantime, several interesting research topics are left for future work. For example, providing theoretical analysis on the optimal choice of kernel hyper-parameters and support recovery for

variable selection is of future research interest. Additionally, it holds great significance in developing more efficient algorithms for variable selection when working with different types of kernels.

Notes

[1]Here we take $\bar{\lambda} = \frac{\mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)}{\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)}$ to satisfy the desired result. Specifically, we provide closed-form expressions on those statistics in the following (see the proof in Appendix EC.2):

$$
\begin{aligned}
A &\triangleq \mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1) = \sqrt{\frac{\tau_1^2}{\tau_1^2+2}} + \sqrt{\frac{\tau_1^2}{\tau_1^2+2(1+\epsilon)^2}} - 2\sqrt{\frac{\tau_1^2}{\tau_1^2+1+(1+\epsilon)^2}},\\
B &\triangleq \sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1) = 4C - 4A^2,\\
C &\triangleq \sqrt{\frac{\tau_1^4}{(\tau_1^2+1)(3+\tau_1^2)}} + \sqrt{\frac{4\tau_1^4}{(\tau_1^2+2)(\tau_1^2+2(1+\epsilon)^2)}}\\
&\quad - \sqrt{\frac{16\tau_1^4}{2\tau_1^2+1+(1+\epsilon)^2+(1+\tau_1^2)((1+\epsilon)^+\tau_1^2)}} - \sqrt{\frac{16\tau_1^4}{(\tau_1^2+1+(1+\epsilon)^2)(\tau_1^2+2(1+\epsilon)^2)}}\\
&\quad + \sqrt{\frac{16\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+(1+\epsilon)^2+2)}} + \sqrt{\frac{\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+3(1+\epsilon)^2)}}.
\end{aligned}
$$

References

[1] Adachi S, Iwata S, Nakatsukasa Y, Takeda A (2017) Solving the trust-region subproblem by a generalized eigenvalue problem. *SIAM Journal on Optimization* 27(1):269–291.

[2] Alizadeh F (1995) Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization* 5(1):13–51.

[3] Bajwa WU, Mixon DG (2012) Group model selection using marginal correlations: The good, the bad and the ugly. *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 494–501 (IEEE).

[4] Beale EML, Tomlin JA (1970) Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *Operations Research* 69(447-454):99.

[5] Berk L, Bertsimas D (2019) Certifiably optimal sparse principal component analysis. *Mathematical Programming Computation* 11:381–420.

[6] Berlinet A, Thomas-Agnan C (2011) *Reproducing kernel Hilbert spaces in probability and statistics* (Springer Science & Business Media).

[7] Bertsimas D, Cory-Wright R, Pauphilet J (2022) Solving large-scale sparse pca to certifiable (near) optimality. *J. Mach. Learn. Res.* 23:13–1.

[8] Bertsimas D, Pauphilet J, Van Parys B (2021) Sparse classification: a scalable discrete optimization perspective. *Machine Learning* 110:3177–3209.

[9] Bonami P, Biegler LT, Conn AR, Cornuéjols G, Grossmann IE, Laird CD, Lee J, Lodi A, Margot F, Sawaya N, et al. (2008) An algorithmic framework for convex mixed integer nonlinear programs. *Discrete optimization* 5(2):186–204.

[10] Bonferroni C (1936) Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8:3–62.

[11] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J, et al. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3(1):1–122.

[12] Boyd SP, Vandenberghe L (2004) *Convex optimization* (Cambridge university press).

[13] Chan SO, Papailliopoulos D, Rubinstein A (2016) On the approximability of sparse pca. *Conference on Learning Theory*, 623–646 (PMLR).

[14] Cheng X, Cloninger A (2022) Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory* 68(10):6631–6662.

[15] Chwialkowski KP, Ramdas A, Sejdinovic D, Gretton A (2015) Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, volume 28.

[16] Dey SS, Kocuk B, Santana A (2019) Convexifications of rank-one-based substructures in qcqps and applications to the pooling problem. *Journal of Global Optimization* 77(2):227–272.

[17] Dey SS, Mazumder R, Wang G (2022) Using ℓ_1-relaxation and integer programming to obtain dual bounds for sparse pca. *Operations Research* 70(3):1914–1932.

[18] Dey SS, Molinaro M, Wang G (2022) Solving sparse principal component analysis with global support. *Mathematical Programming* 1–39.

[19] Dominici* DE (2003) The inverse of the cumulative standard normal probability function. *Integral Transforms and Special Functions* 14(4):281–292.

[20] Fletcher R, Leyffer S (1994) Solving mixed integer nonlinear programs by outer approximation. *Mathematical programming* 66:327–349.

[21] Fukumizu K, Gretton A, Lanckriet G, Schölkopf B, Sriperumbudur BK (2009) Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in neural information processing systems* 22.

[22] Gally T, Pfetsch ME (2016) Computing restricted isometry constants via mixed-integer semidefinite programming. *preprint, submitted* .

[23] Gao R, Kleywegt A (2022) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* .

[24] Good P (2013) *Permutation tests: a practical guide to resampling methods for testing hypotheses* (Springer Science & Business Media).

[25] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. *Journal of Machine Learning Research* 13:723–773.

[26] Gretton A, Fukumizu K, Harchaoui Z, Sriperumbudur BK (2009) A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, volume 22.

[27] Gretton A, Sejdinovic D, Strathmann H, Balakrishnan S, Pontil M, Fukumizu K, Sriperumbudur BK (2012) Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 1205–1213.

[28] Hara S, Morimura T, Takahashi T, Yanagisawa H, Suzuki T (2015) A consistent method for graph based anomaly localization. *Artificial intelligence and statistics*, 333–341 (PMLR).

[29] Idé T, Lozano AC, Abe N, Liu Y (2009) Proximity-based anomaly detection using sparse structure learning. *Proceedings of the 2009 SIAM international conference on data mining*, 97–108 (SIAM).

[30] Idé T, Papadimitriou S, Vlachos M (2007) Computing correlation anomaly scores using stochastic nearest neighbors. *Seventh IEEE international conference on data mining (ICDM 2007)*, 523–528 (IEEE).

[31] Jitkrittum W, Szabó Z, Chwialkowski K, Gretton A (2016) Interpretable distribution features with maximum testing power. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 181?189, NIPS'16.

[32] Kirchler M, Khorasani S, Kloft M, Lippert C (2020) Two-sample testing using deep learning. *International Conference on Artificial Intelligence and Statistics*, 1387–1398 (PMLR).

[33] Kübler JM, Jitkrittum W, Schölkopf B, Muandet K (2022) A witness two-sample test. *International Conference on Artificial Intelligence and Statistics*, 1403–1419 (PMLR).

[34] Kübler JM, Stimper V, Buchholz S, Muandet K, Schölkopf B (2022) Automl two-sample test. *Advances in Neural Information Processing Systems* 35:15929–15941.

[35] Li Y, Fampa M, Lee J, Qiu F, Xie W, Yao R (2022) D-optimal data fusion: Exact and approximation algorithms. *arXiv preprint arXiv:2208.03589* .

[36] Li Y, Xie W (2020) Best principal submatrix selection for the maximum entropy sampling problem: scalable algorithms and performance guarantees. *arXiv preprint arXiv:2001.08537* .

[37] Li Y, Xie W (2020) Exact and approximation algorithms for sparse pca. *arXiv preprint arXiv:2008.12438* .

[38] Li Y, Xie W (2021) Beyond symmetry: Best submatrix selection for the sparse truncated svd. *arXiv preprint arXiv:2105.03179* .

[39] Li Y, Xie W (2022) On the exactness of dantzig-wolfe relaxation for rank constrained optimization problems. *arXiv preprint arXiv:2210.16191* .

[40] Liu F, Xu W, Lu J, Zhang G, Gretton A, Sutherland DJ (2020) Learning deep kernels for non-parametric two-sample tests. *International Conference on Machine Learning*, 6316–6326.

[41] Lofberg J (2004) Yalmip: A toolbox for modeling and optimization in matlab. *2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508)*, 284–289 (IEEE).

[42] Ma S (2013) Alternating direction method of multipliers for sparse principal component analysis. *Journal of the Operations Research Society of China* 1:253–274.

[43] Magdon-Ismail M (2017) Np-hardness and inapproximability of sparse pca. *Information Processing Letters* 126:35–38.

[44] Micchelli CA, Xu Y, Zhang H (2006) Universal kernels. *Journal of Machine Learning Research* 7(12).

[45] Moghaddam B, Weiss Y, Avidan S (2005) Spectral bounds for sparse pca: Exact and greedy algorithms. *Advances in neural information processing systems* 18.

[46] Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B, et al. (2017) Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* 10(1-2):1–141.

[47] Mueller J, Jaakkola T (2015) Principal differences analysis: Interpretable characterization of differences between distributions. *Advances in Neural Information Processing Systems*, volume 28.

[48] Pólik I, Terlaky T (2007) A survey of the s-lemma. *SIAM review* 49(3):371–418.

[49] Schölkopf B, Smola AJ (2002) *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press).

[50] Schrab A, Kim I, Albert M, Laurent B, Guedj B, Gretton A (2021) Mmd aggregated two-sample test. *arXiv preprint arXiv:2110.15073* .

[51] Schrab A, Kim I, Guedj B, Gretton A (2022) Efficient aggregated kernel tests using incomplete u-statistics. *Advances in Neural Information Processing Systems* 35:18793–18807.

[52] Shalev-Shwartz S, Singer Y (2006) Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research* .

[53] Sun Z, Zou S (2021) A data-driven approach to robust hypothesis testing using kernel mmd uncertainty sets. *2021 IEEE International Symposium on Information Theory (ISIT)*, 3056–3061 (IEEE).

[54] Sun Z, Zou S (2022) Kernel robust hypothesis testing. *arXiv preprint arXiv:2203.12777* .

[55] Sutherland DJ, Tung HY, Strathmann H, De S, Ramdas A, Smola A, Gretton A (2016) Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488* .

[56] Taguchi G, Rajesh J (2000) New trends in multivariate diagnosis. *Sankhyā: The Indian Journal of Statistics, Series B* 233–248.

[57] Todd MJ (2001) Semidefinite optimization. *Acta Numerica* 10:515–560.

[58] Van Der Maaten L (2014) Accelerating t-sne using tree-based algorithms. *The journal of machine learning research* 15(1):3221–3245.

[59] Wang J, Chen M, Zhao T, Liao W, Xie Y (2023) A manifold two-sample test study: integral probability metric with neural networks. *Information and Inference: A Journal of the IMA* 12(3):iaad018.

[60] Wang J, Gao R, Xie Y (2021) Two-sample test using projected wasserstein distance. *2021 IEEE International Symposium on Information Theory (ISIT)*, 3320–3325.

[61] Wang J, Gao R, Xie Y (2022) Two-sample test with kernel projected wasserstein distance. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 8022–8055 (PMLR).

[62] Wang J, Moore R, Xie Y, Kamaleswaran R (2022) Improving sepsis prediction model generalization with optimal transport. *Machine Learning for Health*, 474–488 (PMLR).

[63] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al. (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* 347(6226):1138–1142.

Supplementary for *"Variable Selection for Kernel Two-Sample Tests"*

EC.1. ADMM for Solving SDP Problem (15)

Define the domain sets

$$\mathcal{C} = \Big\{ Z \in \mathbb{S}^+_{D+1} :\ Z^{(0,0)} = 1, \mathrm{Tr}(Z) = 2 \Big\},$$

$$\mathcal{B} = \left\{ (Z,q) :\ \sum_{j\in[D]} \left(Z^{(i,j)}\right)^2 \le Z^{(i,i)} q^{(i)}, \left(\sum_{j\in[D]} |Z^{(i,j)}| \right)^2 \le d Z^{(i,i)} q^{(i)}, \forall i \in [D], q \in \overline{\mathcal{Q}} \right\}.$$

Let $\mathcal{I}_{\mathcal{A}}(\cdot)$ denote the indicator function of set $\mathcal{A}$. Then problem (15) can be reformulated as

$$\min_{Z,q} \ \Big\{ -\langle \tilde{A}, Z\rangle + \mathcal{I}_{\mathcal{C}}(Z) + \mathcal{I}_{\mathcal{B}}(Z,q) \Big\}.$$

By introducing a new variable Y, the problem above can be written as

$$\min_{Z,Y,q} \ \Big\{ -\langle \tilde{A}, Z\rangle + \mathcal{I}_{\mathcal{C}}(Z) + \mathcal{I}_{\mathcal{B}}(Y,q) :\ Z = Y \Big\}. \tag{EC.1}$$

The augmented Lagrangian function for problem (EC.1) is defined as

$$\mathcal{L}_\mu(Z,Y,q;\Lambda) = -\langle \tilde{A}, Z\rangle + \mathcal{I}_{\mathcal{C}}(Z) + \mathcal{I}_{\mathcal{B}}(Y,q) - \langle \Lambda, Z - Y\rangle + \frac{1}{2\tau}\|Z-Y\|_F^2,$$

where $\tau > 0$ is a penalty parameter. The ADMM approach produces the following iterations:

$$Z_{k+1} = \arg\min_{Z} \ \mathcal{L}_\mu(Z, Y_k, q_k; \Lambda_k), \tag{EC.2a}$$

$$(Y_{k+1}, q_{k+1}) = \arg\min_{Y,q} \ \mathcal{L}_\mu(Z_{k+1}, Y, q; \Lambda_k), \tag{EC.2b}$$

$$\Lambda_{k+1} = \Lambda_k - \frac{1}{\tau}\big[Z_{k+1} - Y_{k+1}\big]. \tag{EC.2c}$$

The ADMM algorithm terminates at iteration k if for some tolerance parameter $\texttt{tol} > 0$, it holds that

$$\frac{\|Z_{k+1} - Y_{k+1}\|}{1 + \|\tilde{A}\|_1} \le \texttt{tol}.$$

The advantage of ADMM is that, based on the variable splitting trick, the subproblems (EC.2a) and (EC.2b) are easier to solve than the original SDP problem.

Specifically, the subproblem (EC.2a) reduces to

$$Z_{k+1} = \arg\min_{Z\in\mathcal{C}} \ \left\| Z - (Y_k + \tau\tilde{A} + \tau\Lambda_k) \right\|_F^2, \tag{EC.3}$$

which amounts to solving an eigenvalue problem. See the detailed algorithm design in Remark EC.1.

Next, the subproblem (EC.2b) reduces to

$$(Y_{k+1}, q_{k+1}) = \arg\min_{(Y,q)\in\mathcal{B}} \ \|Y - (Z_{k+1} - \tau\Lambda_k)\|_F^2, \tag{EC.4}$$

which amounts to solving a large-scale second-order cone program. See the detailed algorithm design in Remark EC.2.

REMARK EC.1. Given a symmetric matrix $X \in \mathbb{R}^{(D+1)\times(D+1)}$, consider the optimization problem

$$\min_{Z\in\mathcal{C}} \|Z - X\|_F^2.$$

Since this problem is unitary-invariant, its optimal solution is given by $Z^* = U\,\mathrm{diag}(a^*)U^{\mathrm{T}}$ for some vector $a^* \in \mathbb{R}^{D+1}$, where the matrix X admits eigendecomposition $X = U\,\mathrm{diag}(b)U^{\mathrm{T}}$. The vector a^* can be obtained by solving the following problem:

$$a^* = \arg\min \left\{ \|a-b\|_2^2 : \quad a \geq 0, a^{(0)} = 1, \sum_{i=0}^{D} a^{(i)} = 2 \right\}. \tag{EC.5}$$

Such a problem is a variant of the projection problem onto the simplex in Euclidean space. We adopt the algorithm in [52] with complexity $O(D\log D)$ to solve this problem. See details in Algorithm 4.

Algorithm 4 An $O(D\log D)$-complexity algorithm to solving problem (EC.5)

1: Sort $b^{(1:D)}$ to $\hat{b}$ such that $\hat{b}^{(1)} \leq \cdots \leq \hat{b}^{(D)}$.
2: Find smallest index $\hat{j}$ such that $\hat{b}^{(j)} - \frac{1}{D-j+1}\left(\sum_{i=j}^{D}\hat{b}^{(i)} - 1\right) > 0$.
3: Compute $\theta = \frac{1}{D-\hat{j}+1}\left(\sum_{i=\hat{j}}^{D}\hat{b}^{(i)} - 1\right)$
4: **Return** vector a such that $a^{(0)} = 1$ and $a^{(i)} = \max\{0, b^{(i)} - \theta\}, i \in [D]$.

REMARK EC.2. Given a matrix $X \in \mathbb{R}^{(D+1)\times(D+1)}$, consider the optimization problem

$$\min_{(Y,q)\in\mathcal{B}} \|Y - X\|_F^2.$$

It can be reformulated as a second-order cone program that could be solved efficiently based on some off-the-shelf solver:

$$\begin{aligned}
\min_{Y,q\in\overline{\mathcal{Q}},A_i,i\in[D]} \quad & \|\mathrm{vec}(Y) - \mathrm{vec}(X)\|_2^2 \\
\text{s.t.} \quad & \|Y^{(i,:)}\|_1 \leq A_i, i \in [D], \\
& (2A_i, Y^{(i,i)} - dq^{(i)}, Y^{(i,i)} + dq^{(i)}) \in \mathcal{C}_3, Y^{(i,i)} \geq 0, i \in [D], \\
& (Y^{(i,1)}, \ldots, Y^{(i,i)} - \frac{1}{2}q^{(i)}, \ldots, Y^{(i,D)}, \frac{1}{2}q^{(i)}) \in \mathcal{C}_{D+1}, i \in [D],
\end{aligned}$$

where $\mathcal{C}_{D+1}$ denotes the second-order cone of dimension $D+1$:

$$\mathcal{C}_{D+1} = \{(x,t) : \ x \in \mathbb{R}^D, t \in \mathbb{R}, \|x\|_2 \leq t\}.$$

EC.2. Proof of Example 1

Proof of Example 1. Note that the population version of the objective in (7) becomes

$$F(z) = \mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1) - \lambda\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1), \quad \text{if } z^{(1)} \neq 0,$$

when $z = \widehat{z}$ and otherwise $F(z) = 0$. Therefore, taking the variance regularization

$$\lambda \in \left[0, \frac{\mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)}{\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2)}\right),$$

achieves the desired result. It remains to compute $\mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)$ and $\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2)$ to finish the proof. According to the definition, it holds that

$$\begin{aligned}
&\mathrm{MMD}^2(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1)\\
=&\mathbb{E}_{x,x'\sim\mathcal{N}(0,1)}\left[k_1(x,x')\right] + \mathbb{E}_{y,y'\sim\mathcal{N}(0,(1+\epsilon)^2)}\left[k_1(y,y')\right] - 2\mathbb{E}_{x\sim\mathcal{N}(0,1),y\sim\mathcal{N}(0,(1+\epsilon)^2)}\left[k_1(x,y)\right]\\
=&\mathbb{E}_{x,x'\sim\mathcal{N}(0,1)}\left[k_1(x,x') + k_1((1+\epsilon)x,(1+\epsilon)x') - 2k_1(x,(1+\epsilon)y)\right]\\
=&\sqrt{\frac{\tau_1^2}{\tau_1^2+2}} + \sqrt{\frac{\tau_1^2}{\tau_1^2+2(1+\epsilon)^2}} - 2\sqrt{\frac{\tau_1^2}{\tau_1^2+1+(1+\epsilon)^2}},
\end{aligned}$$

where the last step is by substituting the expression $k_1(x,y) = e^{-(x-y)^2/(2\tau_1^2)}$ and calculating several integral of exponential functions. Also, we have that

$$\sigma^2_{\mathcal{H}_1}(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1) = 4\mathbb{E}[H_{1,2}H_{1,3}] - 4\mathrm{MMD}^4(\mathcal{N}(0,1),\mathcal{N}(0,(1+\epsilon)^2);k_1).$$

According to the definition of $H_{i,j}$ in (2), it holds that

$$\begin{aligned}
\mathbb{E}[H_{1,2}H_{1,3}] = \mathbb{E}_{x_1,x_2,x_3,x_4\sim\mathcal{N}(0,1)}\Big[&k_1(x_1,x_2)k_1(x_1,x_3) + 2k_1(x_1,x_2)k_1((1+\epsilon)x_3,(1+\epsilon)x_4)\\
&- 4k_1(x_1,x_2)k(x_1,(1+\epsilon)x_3) - 4k_1(x_1,(1+\epsilon)x_2)k_1((1+\epsilon)x_3,(1+\epsilon)x_4)\\
&+ 4k_1(x_1,(1+\epsilon)x_2)k_1(x_1,(1+\epsilon)x_3) + k_1((1+\epsilon)x_1,(1+\epsilon)x_2)k_1((1+\epsilon)x_1,(1+\epsilon)x_3)\Big]\\
= &\sqrt{\frac{\tau_1^4}{(\tau_1^2+1)(3+\tau_1^2)}} + \sqrt{\frac{4\tau_1^4}{(\tau_1^2+2)(\tau_1^2+2(1+\epsilon)^2)}} - \sqrt{\frac{16\tau_1^4}{2\tau_1^2+1+(1+\epsilon)^2+(1+\tau_1^2)((1+\epsilon)^+\tau_1^2)}}\\
&- \sqrt{\frac{16\tau_1^4}{(\tau_1^2+1+(1+\epsilon)^2)(\tau_1^2+2(1+\epsilon)^2)}} + \sqrt{\frac{16\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+(1+\epsilon)^2+2)}}\\
&+ \sqrt{\frac{\tau_1^4}{(\tau_1^2+(1+\epsilon)^2)(\tau_1^2+3(1+\epsilon)^2)}}
\end{aligned}$$

The proof is completed. □

EC.3. Proofs of Technical Results in Section 3

Proof of Theorem 1 A natural combinatorial reformulation of (STRS) is

$$\max_{\substack{S\subseteq[D]:\ |S|\le d\\ z\in\mathbb{R}^D}} \left\{z^{\mathrm{T}}Az+z^{\mathrm{T}}t:\ \|z\|_2=1, z^{(k)}=0, \forall k\notin S\right\}. \tag{EC.6}$$

Given a size-d set $S\subseteq[D]$, and problem parameters $A\in\mathbb{S}_D, t\in\mathbb{R}^D$, it holds that

$$\begin{aligned}&\max_{z\in\mathbb{R}^D} \left\{z^{\mathrm{T}}Az+z^{\mathrm{T}}t:\ \|z\|_2=1, z^{(k)}=0, \forall k\notin S\right\}\\ =&\max_{z\in\mathbb{R}^d} \left\{z^{\mathrm{T}}A^{(S,S)}z+z^{\mathrm{T}}t^{(S)}:\ \|z\|_2=1\right\}.\end{aligned} \tag{EC.7}$$

Next, we linearize the problem (EC.7) using the auxiliary variable defined as

$$Z=\begin{pmatrix}1\\ z\end{pmatrix}\begin{pmatrix}1\\ z\end{pmatrix}^{\mathrm{T}}=\begin{pmatrix}1 & z^{\mathrm{T}}\\ z & zz^{\mathrm{T}}\end{pmatrix}$$

and the matrix

$$\tilde{A}^{(S,S)}=\begin{pmatrix}0 & \frac{1}{2}(t^{(S)})^{\mathrm{T}}\\ \frac{1}{2}t^{(S)} & A^{(S,S)}\end{pmatrix}.$$

Assume the index of Z and $\tilde{A}^{(S,S)}$ is over $\{0,1,\ldots,d\}^2$. Then we equivalently reformulate the problem (EC.7) as

$$\begin{aligned}\max_{Z\in\mathbb{S}^+_{d+1}}\quad & \langle\tilde{A}^{(S,S)},Z\rangle\\ \text{s.t.}\quad & \mathrm{rank}(Z)=1,\\ & Z^{(0,0)}=1, \mathrm{Tr}(Z)=2.\end{aligned} \tag{EC.8}$$

In particular, constraints $Z\succeq 0, \mathrm{rank}(Z)=1, Z^{(0,0)}=1$ together imply that

$$Z=\begin{pmatrix}1 & z^{\mathrm{T}}\\ z & zz^{\mathrm{T}}\end{pmatrix}$$

for some vector $z\in\mathbb{R}^d$, and the condition $\mathrm{Tr}(Z)=2$ implies $\|z\|_2=1$. By [39, Corollary 3], we further obtain the following equivalent reformulation of problem (EC.7) when dropping the nonconvex rank constraint $\mathrm{rank}(Z)=1$:

$$\begin{aligned}\max_{Z\in\mathbb{S}^+_{d+1}}\quad & \langle\tilde{A}^{(S,S)},Z\rangle\\ \text{s.t.}\quad & Z^{(0,0)}=1, \mathrm{Tr}(Z)=2.\end{aligned} \tag{EC.9}$$

In summary, we obtain the following reformulation of (STRS):

$$\begin{aligned}\max_{Z\in\mathbb{S}^+_{d+1}, S\subseteq[D]:\ |S|\le d}\quad & \langle\tilde{A}^{(S,S)},Z\rangle\\ \text{s.t.}\quad & Z^{(0,0)}=1, \mathrm{Tr}(Z)=2.\end{aligned} \tag{EC.10}$$

It remains to show the equivalence between formulations (10) and (EC.10). We only need to show for any feasible $q\in\mathcal{Q}$ with its support $S:=\{k:\ q^{(k)}=1\}$, it holds that

$$\begin{aligned}&\max_{Z\in\mathbb{S}^+_{D+1}} \left\{\langle\tilde{A},Z\rangle:\ Z_{i,i}\le q^{(i)}, i\in[D], Z^{(0,0)}=1, \mathrm{Tr}(Z)=2\right\}\\ =&\max_{Z\in\mathbb{S}^+_{d+1}} \left\{\langle\tilde{A}^{(S,S)},Z\rangle:\ Z_{0,0}=1, \mathrm{Tr}(Z)=2\right\}.\end{aligned} \tag{EC.11}$$

Since $Z \in \mathbb{S}^+_{D+1}$ is a positive semi-definite matrix, the condition $Z^{(i,i)} = 0$ for $i \in [D] \setminus S$ implies

$$Z^{(i,j)} = 0, \quad \forall (i,j) \notin S \times S.$$

Leveraging this property, we check the relation (EC.11) indeed holds true. □

Proof of Corollary 1. It suffices to verify the following two valid inequalities hold for problem (10):

$$\sum_{j\in[D]} (Z^{(i,j)})^2 \le Z^{(i,i)} q^{(i)}, \quad \forall i \in [D], \tag{EC.12}$$

$$\left(\sum_{j\in[D]} |Z^{(i,j)}| \right)^2 \le d Z^{(i,i)} q^{(i)}, \quad \forall i \in [D]. \tag{EC.13}$$

This verification step follows a similar argument in [37, Lemma 2]. □

Proof of Theorem 2. We first re-write $f(q)$ as the optimal value to the following optimization problem:

$$\begin{aligned}
\max_{Z\in\mathbb{S}^+_{D+1}, U\ge 0, Y, y, t} \quad & \langle \tilde{A}, Z \rangle & \\
& Z^{(0,0)} = 1, & [\lambda_0] \\
& \mathrm{Tr}(Z) = 2, & [\lambda] \\
& \sum_j U^{(i,j)} \le y^{(i)}, \quad \forall i \in [D], & [\beta^{(i)}] \\
& -U^{(i,j)} \le Z^{(i,j)} \le U^{(i,j)}, \quad \forall i,j \in [D], & [W_1^{(i,j)}, W_2^{(i,j)}] \\
& \|(y_i; t_i)\|_2 \le \frac{1}{2} Z^{(i,i)} + \frac{d}{2} q^{(i)}, \quad \forall i \in [D], & [\nu_1^{(i)}] \\
& t_i = \frac{1}{2} Z^{(i,i)} - \frac{d}{2} q^{(i)}, \quad \forall i \in [D], & [\nu_2^{(i)}] \\
& Y^{(i,:)} = Z^{(i,:)} - \frac{1}{2} q^{(i)} e_i, \quad \forall i \in [D], & [\Lambda^{(i,:)}] \\
& \left\| Y^{(i,:)} \right\|_2 \le \frac{1}{2} q^{(i)}, \quad \forall i \in [D]. & [\mu^{(i)}]
\end{aligned}$$

Here, we associate dual variables with primal constraints in brackets. In detail, constraints corresponding to $[\beta^{(i)}], [W_1^{(i,j)}, W_2^{(i,j)}], [\nu_1^{(i)}], [\nu_2^{(i)}]$ are reformulation of the valid inequality (EC.13), and constraints corresponding to $[\Lambda^{(i,:)}]$ and $[\mu^{(i)}]$ are second-order conic reformulation of the valid inequality (EC.12).

Its Lagrangian dual reformulation becomes

$$\begin{aligned}
\min_{\substack{\lambda,\lambda_0,\nu_2,\Lambda \\ \beta,W_1,W_2,\nu_1,\mu\ge 0}} \max_{Z\in\mathbb{S}^+_{D+1}, U\ge 0, Y, y, t} \quad & \langle \tilde{A}, Z \rangle + \lambda_0 \left(1 - Z^{(0,0)}\right) + \lambda \left(2 - \mathrm{Tr}(Z)\right) + \sum_i \beta^{(i)} \left[y^{(i)} - \sum_j U^{(i,j)} \right] \\
& + \sum_{i,j} W_1^{(i,j)} \left[U^{(i,j)} + Z^{(i,j)} \right] + \sum_{i,j} W_2^{(i,j)} \left[U^{(i,j)} - Z^{(i,j)} \right] + \sum_i \nu_1^{(i)} \left[\frac{1}{2} Z^{(i,i)} + \frac{d}{2} q^{(i)} - \|(y_i; t_i)\|_2 \right] \\
& + \sum_i \nu_2^{(i)} \left(\frac{1}{2} Z^{(i,i)} - \frac{d}{2} q^{(i)} - t_i \right) + \sum_i \Lambda^{(i,:)} \left[Z^{(i,:)} - \frac{1}{2} q^{(i)} e_i - Y^{(i,:)} \right] + \sum_i \mu^{(i)} \left(\frac{1}{2} q^{(i)} - \left\| Y^{(i,:)} \right\|_2 \right).
\end{aligned}$$

Or equivalently, it can be written as

$$\min_{\substack{\lambda,\lambda_0,\nu_2,\Lambda \\ \beta,W_1,W_2,\nu_1,\mu\ge 0}} \left\{ \lambda_0 + 2\lambda + \frac{d}{2} (\nu_1 - \nu_2)^{\mathrm{T}} q + \frac{1}{2} (\mu - \mathrm{diag}(\Lambda))^{\mathrm{T}} q \right.$$

$$+\max_{Z\in\mathbb{S}^+_{D+1}}\left\{\langle\tilde{A},Z\rangle-\lambda_0 Z^{(0,0)}-\lambda\mathrm{Tr}(Z)+\sum_{i,j}(W_1^{(i,j)}-W_2^{(i,j)}+\Lambda^{(i,j)})Z^{(i,j)}+\frac{1}{2}\sum_i(\nu_1^{(i)}+\nu_2^{(i)})Z^{(i,i)}\right\}$$
$$+\max_{U\geq 0}\left\{-\sum_i\beta^{(i)}\sum_j U^{(i,j)}+\sum_{i,j}(W_1^{(i,j)}+W_2^{(i,j)})U^{(i,j)}\right\}+\max_Y\left\{-\sum_i\Lambda^{(i,:)}Y^{(i,:)}-\sum_i\mu^{(i)}\|Y^{(i,:)}\|_2\right\}$$
$$+\max_{y,t}\left\{\sum_i\beta^{(i)}y^{(i)}-\sum_i\nu_1^{(i)}\|(y_i;t_i)\|_2-\sum_i\nu_2^{(i)}t_i\right\}\Bigg\}.$$

The inner maximization over Z can be simplified into the constraint

$$\begin{pmatrix}-\lambda_0 & \frac{1}{2}t^{\mathrm{T}}\\ \frac{1}{2}t & A-\lambda I_D+W_1-W_2+\Lambda+\frac{1}{2}\mathrm{diag}(\nu_1+\nu_2)\end{pmatrix}\preceq 0.$$

The inner maximization over U can be simplified as

$$W_1+W_2-\mathrm{diag}(\beta)\leq 0.$$

The inner maximization over Y can be simplified as

$$\sum_j(\Lambda^{(i,j)})^2\leq(\mu^{(i)})^2,\quad i\in[D].$$

The inner maximization over (y,t) can be simplified as

$$(\beta^{(i)})^2+(\nu_2^{(i)})^2\leq(\nu_1^{(i)})^2,\quad i\in[D].$$

Combining those relations, we arrive at the dual problem

$$\begin{aligned}\min_{\substack{\lambda,\lambda_0,\nu_2,\Lambda\\ \beta,W_1,W_2,\nu_1,\mu\geq 0}}\quad & \lambda_0+2\lambda+q^{\mathrm{T}}\left[\frac{d}{2}(\nu_1-\nu_2)+\frac{1}{2}(\mu-\mathrm{diag}(\Lambda))\right]\\ & \begin{pmatrix}-\lambda_0 & \frac{1}{2}t^{\mathrm{T}}\\ \frac{1}{2}t & A-\lambda I_D+W_1-W_2+\Lambda+\frac{1}{2}\mathrm{diag}(\nu_1+\nu_2)\end{pmatrix}\preceq 0,\\ & W_1+W_2-\mathrm{diag}(\beta)\leq 0,\\ & \sum_j(\Lambda^{(i,j)})^2\leq(\mu^{(i)})^2,\quad i\in[D],\\ & (\beta^{(i)})^2+(\nu_2^{(i)})^2\leq(\nu_1^{(i)})^2,\quad i\in[D].\end{aligned}$$

□

Proof of Theorem 3 The left-hand-side relation is easy to show. The proof for the right-hand-side relation is separated into two parts:

- optval(15) $\leq\|t\|_2+d\cdot\{$optval(10) $-\min_k|t[k]|\}$;
- optval(15) $\leq\|t\|_2+D/d\cdot$optval(10).

(I) For any feasible solution (q,Z) to (15), we find

$$\begin{aligned}\sum_i t^{(i)}Z^{(0,i)}&\leq\sum_i|t^{(i)}||Z^{(0,i)}|\leq\sum_i|t^{(i)}|\sqrt{Z^{(0,0)}Z^{(i,i)}}\\ &=\sum_i|t^{(i)}|\sqrt{Z^{(i,i)}}\leq\left(\sum_i|t^{(i)}|^2\right)^{1/2}\left(\sum_i Z^{(i,i)}\right)^{1/2}=\|t\|_2,\end{aligned}$$

where the first inequality is due to taking absolute values, the second inequality is because $Z \succeq 0$ and $|Z^{(0,i)}| \leq \sqrt{Z^{(0,0)} Z^{(i,i)}}$, and the last inequality is by the Cauchy-Schwarz inequality.

As a consequence, for any feasible solution (q, Z) to (15), it holds that

$$\langle \tilde{A}, Z \rangle = \sum_{i,j} A^{(i,j)} Z^{(i,j)} + \sum_i t^{(i)} Z^{(0,i)} \leq \sum_{i,j} \left|A^{(i,j)}\right| \left|Z^{(i,j)}\right| + \|t\|_2. \tag{EC.14}$$

On the other hand, it is easy to verify that for any $i \in [D]$, the following is a feasible solution to (10):

$$Z_i = \begin{pmatrix} 1 \\ e_i \end{pmatrix} \begin{pmatrix} 1 \\ e_i \end{pmatrix}^{\mathrm{T}} \quad \text{or} \quad Z_i = \begin{pmatrix} 1 \\ -e_i \end{pmatrix} \begin{pmatrix} 1 \\ -e_i \end{pmatrix}^{\mathrm{T}},$$

where e_i is a basis vector with the i-th element being 1. This yields

$$\mathsf{optval}(10) \geq \max\left\{A^{(i,i)} + t^{(i)}, A^{(i,i)} - t^{(i)}\right\} = A^{(i,i)} + |t^{(i)}|, \quad \forall i \in [D].$$

Therefore, we obtain

$$A^{(i,i)} \leq \mathsf{optval}(10) - |t^{(i)}| \leq \mathsf{optval}(10) - \min_{i \in [D]} |t^{(i)}|,$$

and $|A^{(i,j)}| \leq \sqrt{A^{(i,i)} A^{(j,j)}} \leq \mathsf{optval}(10) - \min_{i \in [D]} |t^{(i)}|$ for any $i, j \in [D]$. Combining this expression with (EC.14) implies that

$$\langle \tilde{A}, Z \rangle \leq \sum_{i,j} \left|Z^{(i,j)}\right| \cdot \left(\max_{i,j} \left|A^{(i,j)}\right|\right) + \|t\|_2 \leq \sum_{i,j} \left|Z^{(i,j)}\right| \cdot \left(\mathsf{optval}(10) - \min_{i \in [D]} |t^{(i)}|\right) + \|t\|_2. \tag{EC.15}$$

Also, because of the valid inequality $\left(\sum_j \left|Z^{(i,j)}\right|\right)^2 \leq d Z^{(i,i)} q^{(i)}$, it holds that

$$\sum_{i,j} \left|Z^{(i,j)}\right| \leq \sqrt{d} \sum_i \sqrt{Z^{(i,i)} q^{(i)}} \leq \sqrt{d} \left(\sum_i Z^{(i,i)}\right)^{1/2} \left(\sum_i q^{(i)}\right)^{1/2} = d.$$

Combining this relation with (EC.15) gives the desired result.

(II) For any feasible solution (Z, q) in (10), we enforce $Z^{(0,i)} = Z^{(i,0)} = 0$ for $i \in [D]$, then the updated solution is still feasible, with the associated objective value

$$\langle Z^{([D],[D])}, A \rangle.$$

Therefore, we obtain the relation

$$\mathsf{optval}(10) \geq \max_{Z \in \mathbb{S}_D^+, q \in \mathcal{Q}} \left\{ \langle Z, A \rangle : \; Z^{(i,i)} \leq q^{(i)}, i \in [D], \mathrm{Tr}(Z) = 1 \right\} \geq d/D \cdot \lambda_{\max}(A), \tag{EC.16}$$

where the last inequality is due to [37, Proposition 2 and proof of Theorem 5].

For any feasible solution (Z, q) in (15), according to Part (I), it holds that $\sum_i t^{(i)} Z^{(0,i)} \leq \|t\|_2$, and therefore

$$\begin{aligned} \langle \tilde{A}, Z \rangle &= \sum_{i,j} A^{(i,j)} Z^{(i,j)} + \sum_i t^{(i)} Z^{(0,i)} \leq \langle A, Z^{([D],[D])} \rangle + \|t\|_2 \\ &\leq \max_{Z \succeq 0, \mathrm{Tr}(Z) = 1} \langle A, Z \rangle + \|t\|_2 = \lambda_{\max}(A) + \|t\|_2. \end{aligned}$$

Combining this relation with (EC.16) gives the desired result.

□

Proof of Theorem 4(I). Let $z_* = \sum_i y^{(i)} e_i$ be the optimal solution of (STRS), where e_i is the i-th basis vector. Then it holds that

$$\begin{aligned}
\mathsf{optval}(\text{STRS}) &= \sum_i y^{(i)} \left[e_i^{\mathrm{T}}(Az_* + t)\right] \\
&\le \sqrt{\sum_i (y^{(i)})^2} \sqrt{\sum_i (e_i^{\mathrm{T}}(Az_* + t))^2} \\
&\le \sqrt{d} \max_i \ e_i^{\mathrm{T}}(Az_* + t) \\
&\le \sqrt{d} \max_i \ \left\{ \max_{z \in \mathcal{Z}} \ e_i^{\mathrm{T}}(Az + t) \right\} \\
&= \sqrt{d} \max_i \ \left\{ e_i^{\mathrm{T}}(A\hat{z}_i + t) \right\},
\end{aligned}$$

where the last equality is because

$$\hat{z}_i = \arg\max_{z \in \mathcal{Z}} \ e_i^{\mathrm{T}}(Az) = \arg\max_{z \in \mathcal{Z}} \ e_i^{\mathrm{T}}(Az + t).$$

Based on the observation above, one can assert that there exists $i \in [D]$ such that

$$\sqrt{d} e_i^{\mathrm{T}}(A\hat{z}_i + t) \ge \mathsf{optval}(\text{STRS}). \tag{EC.17}$$

Next, we provide the lower bound for $V_{(\mathrm{I})}$:

$$\begin{aligned}
V_{(\mathrm{I})} &= \max_i \ \max \left(e_i^{\mathrm{T}} A e_i + e_i^{\mathrm{T}} t, \hat{z}_i^{\mathrm{T}} A \hat{z}_i + \hat{z}_i^{\mathrm{T}} t \right) \\
&\ge \max_i \ \left\{ \max \left(e_i^{\mathrm{T}} A e_i, \hat{z}_i^{\mathrm{T}} A \hat{z}_i \right) + \min \left(e_i^{\mathrm{T}} t, \hat{z}_i^{\mathrm{T}} t \right) \right\} \\
&\ge \max_i \ \left\{ e_i^{\mathrm{T}} A \hat{z}_i + \min \left(e_i^{\mathrm{T}} t, \hat{z}_i^{\mathrm{T}} t \right) \right\} \\
&= \max_i \ \left\{ e_i^{\mathrm{T}} (A\hat{z}_i + t) + \min \left(0, (\hat{z}_i - e_i)^{\mathrm{T}} t \right) \right\} \\
&\ge \frac{1}{\sqrt{d}} \mathsf{optval}(\text{STRS}) + \max_i \ \min \left(0, (\hat{z}_i - e_i)^{\mathrm{T}} t \right) \\
&\ge \frac{1}{\sqrt{d}} \mathsf{optval}(\text{STRS}) - 2\|t\|_{(d+1)},
\end{aligned}$$

where the second inequality is because $A \succeq 0$ and $0 \le (e_i - \hat{z}_i)^{\mathrm{T}} A (e_i - \hat{z}_i) = (e_i^{\mathrm{T}} A e_i + \hat{z}_i^{\mathrm{T}} A \hat{z}_i) - 2 e_i^{\mathrm{T}} A \hat{z}_i$, i.e.,

$$e_i^{\mathrm{T}} A \hat{z}_i \le \frac{1}{2}(e_i^{\mathrm{T}} A e_i + \hat{z}_i^{\mathrm{T}} A \hat{z}_i) \le \max \left(e_i^{\mathrm{T}} A e_i, \hat{z}_i^{\mathrm{T}} A \hat{z}_i \right),$$

the third inequality is due to (EC.17), and the last inequality is because $\hat{z}_i - e_i$ is a $(d+1)$-sparse vector with $\|\hat{z}_i - e_i\|_2 = 2$, and

$$\max_i \ \min \left(0, (\hat{z}_i - e_i)^{\mathrm{T}} t \right) \ge - \max_i \max_{a:\ \|a\|_0 \le d+1, \|a\|_2 \le 2} a^{\mathrm{T}} t = -2\|t\|_{(d+1)}.$$

The proof is completed. □

Proof of Theorem 4(II). By [1, Theorem 1.1], the primal-dual pair (v, λ) of the trust region subproblem satisfies the following:

$$\begin{cases}
(A - \lambda I) v = -t \\
A \preceq \lambda I \\
\|v\|_2 \le 1 \\
\lambda(1 - \|v\|_2) = 0
\end{cases}$$

Let $\bar{z}$ be the d-sparse truncation of v. Then it holds that

$$\begin{aligned} z^{\mathrm{T}}Av + t^{\mathrm{T}}z &= z^{\mathrm{T}}Av + z^{\mathrm{T}}(-A+\lambda I)v \\ &= \lambda z^{\mathrm{T}}v = \lambda z^{\mathrm{T}}\bar{z} = \lambda\|\bar{z}\|_2 \geq \lambda\sqrt{\frac{d}{D}}. \end{aligned}$$

On the other hand,

$$z^{\mathrm{T}}Av + t^{\mathrm{T}}z \leq \sqrt{z^{\mathrm{T}}Az}\sqrt{v^{\mathrm{T}}Av} + t^{\mathrm{T}}x \leq \sqrt{z^{\mathrm{T}}Az}\cdot\left(\lambda - t^{\mathrm{T}}v\right)^{1/2} + t^{\mathrm{T}}z,$$

where the last inequality is because $(A-\lambda I)v = -t$ and therefore

$$v^{\mathrm{T}}Av + t^{\mathrm{T}}v = \lambda\|v\|_2^2 \leq \lambda.$$

By re-arrangement, it holds that

$$\sqrt{\frac{d}{D}}\lambda \leq \sqrt{z^{\mathrm{T}}Az}\cdot\left(\lambda - t^{\mathrm{T}}v\right)^{1/2} + t^{\mathrm{T}}z.$$

Or equivalently, the dual multiplier λ satisfies

$$\frac{d}{D}\lambda^2 - \left[2\sqrt{\frac{d}{D}}z^{\mathrm{T}}t + z^{\mathrm{T}}Az\right]\lambda + (z^{\mathrm{T}}t)^2 + (z^{\mathrm{T}}Az)(v^{\mathrm{T}}t) \leq 0.$$

Consequently,

$$\frac{d}{D}\lambda^2 - \left[2\sqrt{\frac{d}{D}}\|t\|_{(d)} + z^{\mathrm{T}}Az\right]\lambda - (z^{\mathrm{T}}Az)\|t\| \leq 0.$$

The determinant of the quadratic function on the left-hand-side above is non-negative:

$$\Delta := \left[2\sqrt{\frac{d}{D}}\|t\|_{(d)} + z^{\mathrm{T}}Az\right]^2 + \frac{4d}{D}z^{\mathrm{T}}Az\|t\| \geq 0.$$

On the other hand,

$$\sqrt{\Delta} \leq z^{\mathrm{T}}Az + 2\sqrt{\frac{d}{D}}\|t\|_{(d)} + \frac{2d}{D}\|t\|_2.$$

Hence, we find the upper bound of λ:

$$\begin{aligned} \lambda &\leq \frac{2\sqrt{\frac{d}{D}}\|t\|_{(d)} + z^{\mathrm{T}}Az + \sqrt{\Delta}}{2\frac{d}{D}} \\ &\leq \frac{D}{d}z^{\mathrm{T}}Az + \|t\|_2 + \sqrt{\frac{D}{d}}\|t\|_{(d)} \\ &\leq \frac{D}{d}V_{(\mathrm{II})} + \|t\|_2 + \left(\sqrt{\frac{D}{d}} + \frac{D}{d}\right)\|t\|_{(d)}. \end{aligned}$$

This, together with the fact that $\lambda \geq v^{\mathrm{T}}Av + t^{\mathrm{T}}v \geq \mathsf{optval}(\mathrm{STRS})$ completes the proof. □

Proof of Theorem 5. Define the following two sets:

$$T_d := \{z \in \mathbb{R}^D : \ \|z\|_2 \le 1, \|z\|_1 \le \sqrt{d}\},$$
$$S_d := \{z \in \mathbb{R}^D : \ \|z\|_2 \le 1, \|z\|_0 \le d\}.$$

It has been shown in [17, Lemma 1] that there exists a factor $\rho \in (1, 1+\sqrt{d/(d+1)}]$ such that

$$T_d \subseteq \rho \cdot \mathrm{conv}(S_d).$$

It follows that

$$\begin{aligned}\mathsf{optval}(16) \le & \max_{z \in \rho \cdot \mathrm{conv}(S_d)} \{z^{\mathrm{T}} A z + z^{\mathrm{T}} t\} = \max_{z \in \mathrm{conv}(S_d)} \{\rho^2 z^{\mathrm{T}} A z + \rho z^{\mathrm{T}} t\} \\ = & \max_{z \in S_d} \{\rho^2 z^{\mathrm{T}} A z + \rho z^{\mathrm{T}} t\} \le (\rho^2 - \rho)\|t\|_2 + \max_{z \in S_d} \{\rho^2 z^{\mathrm{T}} A z + \rho^2 z^{\mathrm{T}} t\} \\ = & \rho^2 \mathsf{optval}(\mathrm{STRS}) + (\rho^2 - \rho)\|t\|_2.\end{aligned}$$

□

Proof of Proposition 1. The proof of this proposition is a simple extension from [17]. □

EC.4. Proofs of Technical Results in Section 5

Proof of Proposition 2. (I) We first verify the boundness and Lipschitz continuity conditions. Specifically, it holds that

$$|K_z(x,y)| \le \sum_{s\in[D]} |z^{(s)}| \le \max_{z\in\mathcal{Z}} \sum_{s\in[D]} |z^{(s)}| \le \max_{z\in\mathbb{R}^d:\ \|z\|_2=1} \sum_{s\in[d]} |z^{(s)}| \le \sqrt{d},$$

where the first inequality is because $|k_s(x,y)| \le 1$ for any $x, y \in \mathbb{R}$, and the third inequality is because any vector in $\mathcal{Z}$ only has at most d non-zero entries. Next, we find

$$\begin{aligned} |K_z(x,y) - K_{z'}(x,y)| &= \left| \sum_{s\in[D]} (z^{(s)} - (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) \right| \\ &\le \sum_{s\in[D]} |z^{(s)} - (z')^{(s)}| = \|z - z'\|_1 \\ &\le \sqrt{2d}\|z - z'\|_2, \end{aligned}$$

where the first inequality is because $|k_s(x,y)| \le 1$ for any $x, y \in \mathbb{R}, s \in [D]$, the second inequality is because the vector $z - z'$ only has at most $2d$ non-zero entries. The remaining of Part (I) can be proved by noting that

$$\mathrm{MMD}^2(\mu, \nu, K_{\bar{z}}) \le \max_{z\in\mathcal{Z}} \sum_{s\in[D]} z^{(s)} \mathrm{MMD}^2(\mathrm{Proj}_{s\#}\mu, \mathrm{Proj}_{s\#}\nu; k_s)$$

and

$$\max_{z\in\mathcal{Z}} \ \left|\sigma^2_{\mathcal{H}_1}(\mu, \nu; K_z)\right| \le 8d.$$

(II) For quadratic kernel, we find

$$|K_z(x,y)| \le 2\left(\sum_{s\in[D]} z^{(s)} k_s(x^{(s)}, y^{(s)}) \right)^2 + 2c^2 \le 2d + 2c^2,$$

where the first inequality is based on the relation $(a+b)^2 \le 2a^2 + 2b^2$, and the second inequality is because in Part (I) we have shown that $|\sum_{s\in[D]} z^{(s)} k_s(x^{(s)}, y^{(s)})| \le \sqrt{d}$.

Besides, it holds that

$$|K_z(x,y) - K_{z'}(x,y)| = \left| \sum_{s\in[D]} (z^{(s)} - (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) \right| \left| \sum_{s\in[D]} (z^{(s)} + (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) + 2c \right|.$$

Recall the first term on the right-hand-side can be bounded by $\sqrt{2d}\|z - z'\|_2$, and the second term can be upper bounded by a constant:

$$\begin{aligned} &\left| \sum_{s\in[D]} (z^{(s)} + (z')^{(s)}) k_s(x^{(s)}, y^{(s)}) + 2c \right| \\ \le &\sum_{s\in[D]} |z^{(s)} + (z')^{(s)}||k_s(x^{(s)}, y^{(s)})| + 2c \\ \le &\sum_{s\in[D]} |z^{(s)} + (z')^{(s)}| + 2c \\ \le &\max_{v:\ \|v\|_0 \le 2d, \|v\|_2 \le 2} \|v\|_1 + 2c \\ \le &2\sqrt{2d} + 2c. \end{aligned}$$

Combining those two relations gives the desired result. The remaining of Part (II) can be proved by noting that

$$\mathrm{MMD}^2(\mu,\nu,K_{\tilde{z}}) \le \max_{z\in\mathcal{Z}} \max_{z\in\mathcal{Z}} z^{\mathrm{T}}\mathcal{A}(\mu,\nu)z + z^{\mathrm{T}}\mathcal{T}(\mu,\nu)$$

and

$$\max_{z\in\mathcal{Z}} \left|\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \le 8(2d+2c^2)^2.$$

(III) The boundness of Gaussian kernel is easy to check. The Lipscthiz continuity condition of the Gaussian kernel follows from [40, Lemma 20]. The remaining of Part (III) can be proved by noting that

$$\max_{z\in\mathcal{Z}} \left|\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \le 8.$$

Before showing the proof of Theorem 6, we list two useful technical lemmas.

LEMMA EC.1 **([25, Theorem 10])**. *Assume the kernel $K_z(\cdot,\cdot)$ is uniformly bounded, i.e., for any $z \in \mathcal{Z}, x, y \in \Omega$, it holds that $|K_z(x,y)| \le M$. For fixed $z \in \mathcal{Z}$, with probability at least $1-\delta$,*

$$\left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| \le \frac{16M}{\sqrt{2n}}\sqrt{\log\frac{2}{\delta}}.$$

LEMMA EC.2 **([40, Lemma 17 and 18])**. *Assume the kernel $K_z(\cdot,\cdot)$ is uniformly bounded, i.e., for any $z \in \mathcal{Z}, x, y \in \Omega$, it holds that $|K_z(x,y)| \le M$. For fixed $z \in \mathcal{Z}$, with probability at least $1-\delta$,*

$$\left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \le 448\sqrt{\frac{2}{n}\log\frac{2}{\delta}} + \frac{1152M^2}{n}.$$

Proof of Theorem 6 We first consider an ϵ-cover of $\mathcal{Z}$, denoted as $\{z_i\}_{i\in[T]}$. According to the definition of $\mathcal{Z}$, it can be shown that $T \le \binom{D}{d}(4/\epsilon)^d$. Applying the union bound regarding the concentration inequality in Lemma EC.1, we obtain with probability at least $1-\delta$,

$$\max_{z\in\{z_i\}_{i\in[T]}} \left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| \le \frac{16M}{\sqrt{2n}}\sqrt{\log\frac{2T}{\delta}}.$$

For any $z \in \mathcal{Z}$, there exists z' from $\{z_i\}_{i\in[T]}$ such that $\|z-z'\|_2 \le \epsilon$. Based on the Lipschitz assumption regarding the kernel function, we find with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| \\
\le &\max_{z\in\{z_i\}_{i\in[T]}} \left|S^2(\mathbf{x}^n,\mathbf{y}^n;K_z) - \mathrm{MMD}^2(\mu,\nu;K_z)\right| + 8L\epsilon \\
\le &\frac{16M}{\sqrt{2n}}\sqrt{\log\frac{2T}{\delta}} + 8L\epsilon \\
\le &\frac{16M}{\sqrt{2n}}\sqrt{\log\binom{D}{d}\frac{2}{\delta} + d\log\frac{4}{\epsilon}} + 8L\epsilon.
\end{aligned}$$

Setting $\epsilon = 1/\sqrt{n}$ gives the desired result.

Next, applying the union bound regarding the concentration inequality in Lemma EC.2, we obtain with probability at least $1-\delta$,

$$\max_{z\in\{z_i\}_{i\in[T]}} \left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \le 448\sqrt{\frac{2}{n}\log\frac{2T}{\delta}} + \frac{1152M^2}{n}.$$

Similar as in the first part, we find with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| \\
\leq &\max_{z\in\{z_i\}_{i\in[T]}} \left|\hat{\sigma}^2_{\mathcal{H}_1}(\mathbf{x}^n,\mathbf{y}^n;K_z) - \sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z)\right| + 512LM\epsilon \\
\leq &448\sqrt{\frac{2}{n}\log\frac{2T}{\delta}} + \frac{1152M^2}{n} + 512LM\epsilon \\
\leq &448\sqrt{\frac{2}{n}\log\binom{D}{d}\frac{2}{\delta} + \frac{2}{n}d\log\frac{4}{\epsilon}} + \frac{1152M^2}{n} + 512LM\epsilon.
\end{aligned}$$

Also, setting $\epsilon = 1/\sqrt{n}$ gives the desired result.

Proof of Theorem 7. To simplify notation, let us define the population version of the objective in (7) as follows:

$$F^*(z) = \mathrm{MMD}^2(\mu,\nu;K_z) - \lambda\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z).$$

We first derive the lower bound of $F^*(\hat{z}_{\mathrm{Tr}})$ in terms of $F^*(\bar{z})$ with $\bar{z}$ defined in Assumption 2 using concentration analysis. It is clear that $|F^*(z) - F(z)| \leq \epsilon_{n,\delta/2}$ with probability at least $1-\delta$. As a consequence, with probability at least $1-\delta$, it holds that

$$F^*(\hat{z}_{\mathrm{Tr}}) \geq F(\hat{z}_{\mathrm{Tr}}) - \epsilon_{n_{\mathrm{Tr}},\delta/4} \geq F(\bar{z}) - \epsilon_{n_{\mathrm{Tr}},\delta/4} \geq F^*(\bar{z}) - 2\epsilon_{n_{\mathrm{Tr}},\delta/4}, \tag{EC.18}$$

where we use this observation in the first and last inequalities, and the second inequality is because of the sub-optimality of $\bar{z}$. Now we are ready to show part (I) of this theorem. By definition, we find

$$\begin{aligned}
\mathbb{E}[T_{n_{\mathrm{Te}}}] &= \mathrm{MMD}^2(\mu,\nu;K_{\hat{z}_{\mathrm{Tr}}}) \\
&= F^*(\hat{z}_{\mathrm{Tr}}) + \lambda\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_{\hat{z}_{\mathrm{Tr}}}) \geq F^*(\hat{z}_{\mathrm{Tr}}) + \lambda\min_{z\in\mathcal{Z}}\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z).
\end{aligned}$$

Combining the relation above and (EC.18) implies that, with probability at least $1-\delta$, it holds that

$$\mathbb{E}[T_{n_{\mathrm{Te}}}] \geq F^*(\bar{z}) - 2\epsilon_{n_{\mathrm{Tr}},\delta/4} + \lambda\min_{z\in\mathcal{Z}}\sigma^2_{\mathcal{H}_1}(\mu,\nu;K_z) = \Delta_{\bar{z}} - 2\epsilon_{n_{\mathrm{Tr}},\delta/4}.$$

The second part of this theorem follows from [25, Theorem 12]. □

LEMMA EC.3 **(Asymptotics of Inverse Error Function [19]).** *Denote by $S(x)$ the inverse of the error function*

$$\Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,\mathrm{d}t.$$

As $x \to 1$, it holds that

$$S(x) \to \sqrt{\mathcal{LW}\left(\frac{1}{2\pi(x-1)^2}\right)},$$

where $\mathcal{LW}(x)$ denotes the function Lambert $W(x)$ admiting the series expansion

$$\mathcal{LW}(x) = \sum_{n\geq 1}\frac{(-1)^{n-1}}{n!}x^n.$$

Specifically, $\mathcal{LW}(x) \to \ln(x) - \ln\ln(x)$ as $x \to \infty$.

Proof of Theorem 8. It is worth noting that

$$\mathbb{P}(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}}) = \mathbb{P}\left(T_{n_{\mathrm{Te}}} > \frac{\tau}{n_{\mathrm{Te}}}\right) = 1 - \mathbb{P}\left(\frac{\sqrt{n_{\mathrm{Te}}}(T_{n_{\mathrm{Te}}} - \mathbb{E}T_{n_{\mathrm{Te}}})}{\sigma_{\mathcal{H}_1}} \le \frac{\tau}{\sigma_{\mathcal{H}_1}\sqrt{n_{\mathrm{Te}}}} - \frac{\sqrt{n_{\mathrm{Te}}}\mathbb{E}T_{n_{\mathrm{Te}}}}{\sigma_{\mathcal{H}_1}}\right)$$
$$\ge 1 - \Phi\left(\frac{\tau}{\sigma_{\mathcal{H}_1}\sqrt{n_{\mathrm{Te}}}} - \frac{\sqrt{n_{\mathrm{Te}}}\mathbb{E}T_{n_{\mathrm{Te}}}}{\sigma_{\mathcal{H}_1}}\right) - \frac{C\rho}{\sigma_{\mathcal{H}_1}^3\sqrt{n_{\mathrm{Te}}}},$$

where for the inequality above we apply the Berry–Esseen theorem to argue that the distribution of $\sqrt{n_{\mathrm{Te}}}(T_{n_{\mathrm{Te}}} - \mathbb{E}T_{n_{\mathrm{Te}}})/\sigma_{\mathcal{H}_1}$ can be approximated as the normal distribution with residual error $O(1/\sqrt{n_{\mathrm{Te}}})$. Therefore, as long as we ensure that

$$\frac{\tau}{\sigma_{\mathcal{H}_1}\sqrt{n_{\mathrm{Te}}}} - \frac{\sqrt{n_{\mathrm{Te}}}\mathbb{E}T_{n_{\mathrm{Te}}}}{\sigma_{\mathcal{H}_1}} \le \Phi^{-1}(\epsilon) \Longleftrightarrow \mathbb{E}T_{n_{\mathrm{Te}}} \ge \frac{\tau}{n_{\mathrm{Te}}} + \frac{\Phi^{-1}(1-\epsilon)}{\sqrt{n_{\mathrm{Te}}}},$$

it holds that the testing power is lower bounded:

$$\mathbb{P}(T_{n_{\mathrm{Te}}} > t_{\mathrm{thres}}) \ge 1 - \epsilon - \frac{C\rho}{\sigma_{\mathcal{H}_1}^3\sqrt{n_{\mathrm{Te}}}}.$$

Taking $\epsilon = 1/\sqrt{n_{\mathrm{Te}}}$ and applying the asymptotic formula on the inverse cdf $\Phi^{-1}(\cdot)$ in Lemma EC.3 gives the desired result. The type-I risk upper bound follows a similar argument. □