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 features to best distinguish samples from two groups. We propose a kernel maximum mean discrepancy (MMD) framework to solve this problem and further derive its equivalent mixed-integer programming formulations for linear, quadratic, and Gaussian types of kernel functions. Our proposed framework admits advantages of both computational efficiency and nice statistical properties: (i) A closed-form solution is provided for the linear kernel case. Despite NP-hardness, we provide an exact mixed-integer semi-definite programming formulation for the quadratic kernel case, which further motivates the development of exact and approximation algorithms. We propose a convex-concave procedure that finds critical points for the Gaussian kernel case. (ii) We provide non-asymptotic uncertainty quantification of our proposed formulation under null and alternative scenarios. Experimental results demonstrate good performance of our framework.

1. Introduction

We study the variable selection problem for two-sample testing, which aims to select the most informative features to best distinguish samples from two groups. On the one hand, identifying interpretable features that cause the intrinsic difference between populations is an important task for general scientific discovery areas. For example, only a small subset of variables (such as gene expressions, biological indicators, etc) may cause the differences between normal and abnormal data samples [51]. On the other hand, the differences between two high-dimensional datasets usually exhibit low-dimensional structure [47], and therefore extracting a small group of important features as a pre-processing step will improve the power of high-dimensional two-sample testing.

Unfortunately, the selection of key features can be challenging primarily for three reasons: (i) one does not have much information regarding the data-generating distribution for each group; (ii) the number of observed samples is insufficient such that one does not have accurate estimates of distributions for each group; (iii) the dimension of data points is too high such that it is difficult to compare two groups of data.

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 & Rajesh [45] assumes target distributions as Gaussian and finds important features such that the difference between mean and covariance among two groups is maximized. Following this seminal work, references [27, 28, 24] 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 [6] has been proposed in the two-sample testing context to compare every single feature using statistical tests to obtain representative features, but it may not perform well when correlations exist between features.

Recently, Mueller & Jaakkola [39] 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 [18], this approach serves as a non-parametric way for selecting features. However, the non-asymptotic convergence of the proposed projected Wasserstein distance has non-negligible dependence on the size of distribution support, projected dimension, etc. Besides, due to the nonconvexity of this problem, only approximation algorithms are proposed to find local optimum solutions, while in statistical analysis it is assumed that one can succeed in finding the global optimum.

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 [22, 20, 21]. The MMD-based approaches have been proposed in literature to study the problems of one- or two-sample testing [22, 20, 21, 35, 29, 10, 43, 44]. In this paper, we apply MMD to design a novel variable selection framework for two-sample testing. Our contributions are summarized as follows:

(I) We provide an exact solution for the linear kernel case, and show this framework seeks to find the projection direction such that the difference between mean vectors of target distributions is maximized.

(II) We reformulate the framework for the quadratic kernel case as an inhomogeneous quadratic maximization problem with ℓ_2 and ℓ_0 norm constraints, which is a slight extension of the classical sparse PCA problem. 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 literature. Besides, we reveal this framework seeks to find the projection direction to maximize the difference between the combination of first- and second-order moments of target distributions.

(III) We approximate the variable selection problem for the Gaussian kernel case as a nonconvex matrix optimization over a spectrahedron, and develop convex-concave procedure to find critical points. We show the Gaussian kernel MMD framework can detect informative features for arbitrarily two distinct distributions, making it a suitable choice for non-parametric variable selection.

(IV) To quantify the false alarm rate, we provide non-asymptotic uncertainty quantification on our proposed formulations for null and alternative scenarios.

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

Notations Define $\mathbb{F} = \{0, 1\}$ and $\|\cdot\|_{\text{op}}$ as the operator norm. Given a positive integer n, define $[n] = \{1, \ldots, n\}$. Let $\mathbb{S}_n^+$ denote the collection of $n \times n$ symmetric positive semi-definite matrices. Given an $m \times n$ matrix A and two sets $S \subseteq [m], t \subseteq [n]$, denote $A_{S,T}$ as the submatrix with rows and columns indexed by S and T. Given a vector $z \in \mathbb{R}^D$, we use $z[k]$ to denote the k-th entry in z. 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 $x \sim \mu$.

1.1. Related Work

Classical variable selection approaches seek to extract the most valuable features from a group of high-dimensional data points. Especially, the sparse PCA approach aims to select crucial features such that the sample covariance based on sample sets can be maximized [31, 15, 14]; the truncated SVD approach aims to formulate a low-rank data matrix with minimum approximation error [30], and the maximum entropy sampling or experiment design approach aims to select a subgroup of samples that reserve information as much as possible [32, 34]. However, variable selection for identifying the differences between two groups is less studied in literature. Recently, Mueller & Jaakkola [39] 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. [46] and Wang et al. [48] 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. However, these two references do not provide methodologies to identify useful subsets of features that characterize the differences between two groups the most.

2. Model Formulation

Let $\mathrm{x}^n := \{x_i\}_{i=1}^n$ and $\mathrm{y}^m = \{y_i\}_{i=1}^m$ be i.i.d. samples generated from distributions μ and ν, respectively. In particular, those data samples are in the Euclidean space $\mathbb{R}^D$, where the dimension D denotes the number of feature variables. In this paper, we study variable selection based on the maximum mean discrepancy (MMD), which quantifies the discrepancy between two probability distributions using test functions in a reproducing kernel Hilbert space (RKHS).

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

$$\sum_{i,j} c_i c_j K(x_i, x_j) \geq 0$$

for any finite set of samples $\{x_i\}_{i=1}^N$ in $\mathbb{R}^D$ and $\{c_i\}_{i=1}^N$ in $\mathbb{R}$. A positive semi-definite kernel K induces a unique RKHS $\mathcal{H}_K$. Given a RKHS $\mathcal{H}_K$ containing a class of candidate testing functions and two distributions μ, ν, define the corresponding MMD statistic as

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

It can be shown that the MMD statistic can be equivalently written as a distance in $\mathcal{H}$ between mean embeddings [7]:

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

Although the data distributions μ and ν are not available, we can formulate an estimate of $\mathrm{MMD}(\mu, \nu; K)^2$ based on samples x^n and y^m using the following statistic:

$$S^2(\mathrm{x}^n, \mathrm{y}^m; K) \triangleq \frac{1}{n^2} \sum_{i,j \in [n]} K_{i,j}^{x,x} + \frac{1}{m^2} \sum_{i,j \in [m]} K_{i,j}^{y,y} - \frac{2}{mn} \sum_{i \in [n], j \in [m]} K_{i,j}^{x,y}. \tag{1}$$

Here for notational simplicity, we write $K_{i,j}^{x,x}$ for $K(x_i, x_j)$ and define similar notations for $K_{i,j}^{x,y}$ and $K_{i,j}^{y,y}$.

In this paper, we focus on the variable selection task with MMD by considering the following type of kernel function:

$$K_z(x, y) = \psi\left(\sum_{k \in [D]} \phi(x[k], y[k]; z[k]) \right), \tag{2}$$

where $\phi : \mathbb{R}^3 \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ are mappings to be specified, and $z = (z[k])_{k \in [D]}$ is the sparse variable selection coefficient to be optimized. Specially, we aim to pick the optimal projection vector z^* such that the empirical projected MMD statistic is maximized:

$$\begin{aligned} \max_{z \in \mathcal{Z}} \quad & S^2(\mathrm{x}^n, \mathrm{y}^m; K_z) \\ \text{where} \quad & z \in \mathcal{Z} := \{z \in \mathbb{R}^D : \|z\|_2 = 1, \|z\|_0 = d\}. \end{aligned} \tag{3}$$

For $k \in [D]$, the k-th feature of data points is selected if the k-th entry in z^* is non-zero. The goal of (3) is to pick a subset of d features such that the difference between two groups of data points under the selection coefficient z^* is as much as possible.

Algorithm 1 A permutation two-sample test using MMD

Require: cardinality d, number of permutation times N_p, collected samples x^n and y^m.

1: Split data as $\mathrm{x}^n = \mathrm{x}^{\mathrm{Tr}} \cup \mathrm{x}^{\mathrm{Te}}$ and $\mathrm{y}^m = \mathrm{y}^{\mathrm{Tr}} \cup \mathrm{y}^{\mathrm{Te}}$.
2: Obtain $z^* = \arg\max \left\{ S^2(\mathrm{x}^{\mathrm{Tr}}, \mathrm{y}^{\mathrm{Tr}}; K_z) : z \in \mathcal{Z} \right\}$.
3: Compute the statistic $T = S^2(\mathrm{x}^{\mathrm{Te}}, \mathrm{y}^{\mathrm{Te}}; K_{z^*})$.
4: **for** $t = 1, \ldots, N_p$ **do**
5: Shuffle $\mathrm{x}^{\mathrm{Te}} \cup \mathrm{y}^{\mathrm{Te}}$ to obtain $\mathrm{x}^{\mathrm{Te}}_{(t)}$ and $\mathrm{y}^{\mathrm{Te}}_{(t)}$.
6: Compute the statistic for permuted samples:

$$T_t = S^2(\mathrm{x}^{\mathrm{Te}}_{(t)}, \mathrm{y}^{\mathrm{Te}}_{(t)}; K_{z^*}).$$

7: **end for**
Return the p-value $\frac{1}{N_p} \sum_{t=1}^{N_p} 1\{T_t \geq T\}$.

Based on the proposed variable selection framework, we present a kernel two-sample test in the following. We split the data points into training and testing datasets. We first use the training set to obtain the selection coefficient that optimally identifies the difference between two groups. Next, we perform the permutation test on testing data points that are projected based on the trained selection coefficient. The detailed algorithm is presented in Algorithm 1. This test is guaranteed to exactly control the type-I error [19] because we evaluate the p-value of the test via the permutation approach.

The choice of kernel function $K_z(\cdot, \cdot)$ largely influences the performance of variable selection and two-sample test. The key to picking a valid kernel is to make the optimal value of the population counterpart of the formulation (3) non-vanishing. Otherwise, in any direction will the population MMD statistic leads to the same value, indicating the constructed kernel is not powerful enough for the variable selection task. In the following subsection, we will consider a simple linear kernel, demonstrating it may not achieve satisfactory performance for general distributions. In Section 3 and 4, we will discuss more advanced quadratic and Gaussian kernels, respectively. Those three types of kernels are specified in the following:

Linear Kernel: consider $\phi(x, y; z) = zxy$ and $\psi(x) = x$ in (2):

$$K_z(x, y) = \sum_{k \in [D]} z[k]x[k]y[k]. \tag{4}$$

Quadratic Kernel: consider $\phi(x, y; z) = zxy$ and $\psi(x) = (x + c)^2$ in (2), where $c > 0$ is the kernel bandwidth hyper-parameter:

$$K_z(x, y) = \left(\sum_{k \in [D]} z[k]x[k]y[k] + c \right)^2. \tag{5}$$

Gaussian kernel: consider $\phi(x, y; z) = z^2(x - y)^2$ and $\psi(x) = e^{-x/(2\gamma)}$ in (2), where $\gamma > 0$ is the kernel bandwidth hyper-parameter:

$$K_z(x, y) = \exp\left[-\frac{\left(\sum_{k \in [D]} z[k](x[k] - y[k]) \right)^2}{2\gamma} \right]. \tag{6}$$

2.1. Example: Linear Kernel MMD

We first consider the linear kernel case in (4). Based on this expression, we reformulate (1) as a simple mixed-integer linear optimization (MILO) problem:

$$\max_{z \in \mathcal{Z}} \ a^{\mathrm{T}} z, \tag{Linear-MMD}$$

where the k-th entry of the coefficient vector a is

$$a[k] = \left(\frac{1}{n} \sum_{i \in [n]} x_i[k] - \frac{1}{m} \sum_{j \in [m]} y_j[k] \right)^2 .$$

THEOREM 1. *Let $\{k_1, \ldots, k_D\}$ be the sorted indices of $\{1, \ldots, D\}$ such that $a[k_1] \geq \cdots \geq a[k_D]$. An optimal solution to the MILO problem* (Linear-MMD) *is*

$$z^*[k] = a[k] \cdot \left(\sum_{k' \in \{k_1, \ldots, k_d\}} a[k']^2 \right)^{-1/2}$$

if $k \in \{k_1, \ldots, k_d\}$ and otherwise $z^*[k] = 0$.

REMARK 1 (CONS ABOUT LINEAR KERNEL). Under the linear kernel choice, it can be shown that

$$\mathrm{MMD}^2(\mu, \nu; K_z) = \sum_{k \in [D]} z[k] \big(\overline{x}[k] - \overline{y}[k] \big)^2,$$

where $\overline{x}, \overline{y} \in \mathbb{R}^D$ are mean vectors for distributions μ and ν, respectively. In other words, the selection coefficient aims to find a direction such that the difference between the mean of target distributions μ and ν is maximized. Under the case where $\overline{x} = \overline{y}$, the linear kernel MMD does not have enough power to find informative features to distinguish those two distributions.

3. Algorithms for Quadratic MMD

Next, we consider the quadratic kernel choice in (5), which can be equivalently re-written as a quadratic form:

$$K_z(x, y) = z^{\mathrm{T}} A(x, y) z + z^{\mathrm{T}} t(x, y) + c^2.$$

Here we define the vector $a(x, y) := (x[k]y[k])_{k \in [D]}$, the matrix $A(x, y) := a(x, y)a(x, y)^{\mathrm{T}}$, and the vector $t(x, y) := 2ca(x, y)$. Consequently, the problem (3) can be formulated as a nonconvex mixed-integer quadratic program (MIQP):

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

where the coefficients

$$A = \frac{\sum_{i,j \in [n]}^{n} A(x_i, x_j)}{n^2} + \frac{\sum_{i,j \in [m]} A(y_i, y_j)}{m^2} - \frac{2 \sum_{i \in [n], j \in [m]} A(x_i, y_j)}{mn},$$
$$t = \frac{\sum_{i,j \in [n]} t(x_i, x_j)}{n^2} + \frac{\sum_{i,j \in [m]} t(y_i, y_j)}{m^2} - \frac{2 \sum_{i \in [n], j \in [m]} t(x_i, y_j)}{mn}.$$

The objective in (Quad-MMD) can be re-written as $z^{\mathrm{T}}(A - \lambda_{\min}(A)I)z + z^{\mathrm{T}} t + \lambda_{\min}(A)$, where $A - \lambda_{\min}(A)I \succeq 0$. One can therefore assume the matrix A in (Quad-MMD) is positive semi-definite without loss of generality. Specially, when the scalar $c = 0$, the above problem reduces to the sparse PCA formulation, which has been studied extensively in literature [3, 17, 38, 31]. Otherwise, it can be solved exactly based on the branch-and-bound algorithm, which can be implemented using the off-the-shelf solver Gurobi. Unfortunately, there are two concerns about this formulation that limits its application in large-scale scenarios. First, since the objective function is non-concave in z, it is challenging to develop exact algorithms for solving (Quad-MMD). We provide a mixed-integer convex programming formulation which seems to be easier for developing exact algorithms. Second, this problem is NP-hard even if $t = 0$ as pointed out in Magdon-Ismail [37]. In the case of large problem size, we provide its convex relaxation that can be solved more efficiently.

3.1. Equivalent Mixed-Integer Semi-definite Programming (MISDP) Reformulation

We first provide an exact MISDP reformulation of (Quad-MMD). When the coefficient vector $t = 0$, similar reformulation results have been developed in the sparse PCA literature [31, 5]. However, such a reformulation for $t \neq 0$ is new in literature.

THEOREM 2. *Define the matrix*

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

Problem (Quad-MMD) *can be equivalently formulated as*

$$\max_{Z \in \mathbb{S}^+_{D+1},\, q \in \mathcal{Q}} \quad \langle \tilde{A}, Z \rangle \tag{7a}$$

$$s.t. \quad Z_{i,i} \leq q[i], \quad i \in [D], \tag{7b}$$

$$Z_{0,0} = 1, Tr(Z) = 2, \tag{7c}$$

where the set $\mathcal{Q} = \left\{ q \in \mathbb{F}^D : \ \sum_{k \in [D]} q_i = d \right\}$, *and we assume the indices of* $Z, \tilde{A} \in \mathbb{S}^+_{D+1}$ *are both over* $\{0, 1, \ldots, D\}^2$. *Its continuous relaxation value equals*

$$w_{rel} = \max_{z:\ \|z\|_2 = 1} \left\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} t \right\}. \tag{8}$$

The proof idea of Theorem 2 is to express the problem (Quad-MMD) as a *rank*-1 *constrained SDP* problem. Leveraging some well-known results on rank constrained optimization (see, e.g., Pólik & Terlaky [41], Dey et al. [13], Li & Xie [33]), one can remarkably remove the rank constraint without changing the optimal value of the original SDP problem. Although (7) is equivalent to (Quad-MMD), the fact that its continuous relaxation value is equal to w_{rel} suggests that it might be a weak formulation. This motivates us to propose two valid inequalities to strengthen the formulation (7) in Corollay 1.

COROLLARY 1. *The problem* (Quad-MMD) *reduces to the following stronger MISDP formulation:*

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

$$s.t. \quad (7c),$$

$$|Z_{i,j}| \leq M_{i,j} q[i], \quad \forall i, j \in [D], \tag{9a}$$

$$\sum_{j \in [D]} |Z_{i,j}| \leq \sqrt{d} q[i], \quad \forall i \in [D]. \tag{9b}$$

Here the matrix $M \in \mathbb{R}^{D \times D}$ *is defined as*

$$M_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 1/2, & \text{otherwise}. \end{cases} \tag{9c}$$

It is worth noting that similar inequalities have been proposed in literature to improve the performance of solving sparse PCA [31, 14, 5]. On the one hand, the resulting formulation (9) can be directly solved via some exact MISDP solvers such as YALMIP [36]. 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.

Algorithm 2 Exact Algorithm for solving (Quad-MMD)

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

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

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

3.2. Exact Algorithm: Cutting-Plane Method

To develop exact algorithm, we first reformulate the problem (9) as a max-min saddle point problem so that it can be solved based on the outer-approximation technique.

THEOREM 3. *Problem* (7) *shares the same optimal value as the following problem:*

$$\max_{q \in \mathcal{Q}} \quad f(q), \tag{10}$$

where the function $f(q)$ *is concave in* q *and is the optimal value to the following problem:*

$$\min_{\substack{\lambda, \lambda_0 \\ \alpha^+, \alpha^-, \beta \ge 0}} \quad \lambda_0 + 2\lambda_1 + \sum_i q[i] \left\{ \sum_j M_{i,j} \max(0, \alpha^+_{i,j} + \alpha^-_{i,j} - \beta[i]) + \sqrt{d}\beta[i] \right\}$$
$$s.t. \quad \begin{pmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 I_D + \alpha^+ - \alpha^- \end{pmatrix} \succeq \tilde{A}.$$

By Theorem 3, 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 super-gradient of f at $\hat{q}$. Based on this observation, we apply the outer-approximation technique to solve the maximization problem: at iterations $t = 1, 2, \ldots, T$, we maximize and refine a piecewise linear overestimator of $f(q)$:

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

We summarize the proposed algorithm in Algorithm 2.

3.3. Approximation Algorithm: Convex Relaxation

When the problem size is large, solving (7) is intractable in general. In this case, one can solve its convex relaxation formulation:

$$\begin{aligned} \max_{Z \in \mathbb{S}^+_{D+1}, q \in \overline{\mathcal{Q}}} \quad & \langle \tilde{A}, Z \rangle \\ \text{s.t.} \quad & (7c), (9a), (9b), \end{aligned} \tag{11}$$

where the set $\overline{\mathcal{Q}} = \left\{ q \in [0,1]^D : \ \sum_{k \in [D]} q[k] = d \right\}$. After the solution from (11) is obtained, one can obtain a feasible solution in $\mathcal{Q}$ using a greedy algorithm. We summarize the proposed algorithm in Algorithm 3. In the following theorem, we show that the optimal value of the convex relaxation problem is not far from that of the original problem (7). The proof adopts similar technique as in [31, Theorem 5], but we extend the analysis for inhomogeneous quadratic maximization formulation.

Algorithm 3 Approximation Algorithm for solving problem (Quad-MMD)

1: Compute q^* as the optimal solution from (11)
2: Project q^* back to $\mathcal{Q}$.
3: Compute Z as the optimal solution from

$$\max_{Z\in\mathbb{S}^+_{D+1}} \left\{\langle \tilde{A}, Z\rangle :\ Z_{0,0}=1, \mathrm{Tr}(Z)=2, Z_{i,j}=0 \text{ if } q[i]q[j]=0, \forall i,j\in[D]\right\}.$$

4: **Return** (q, Z)

THEOREM 4. *Denote by* optval(11) *and* optval(7) *the optimal values of problem* (11) *and* (7)*, respectively. Then it holds that*

$$\textit{optval}(7) \le \textit{optval}(11) \le \|t\|_2 + \min\left\{D/d\cdot \textit{optval}(7), d\cdot \textit{optval}(7) - \min_k |t[k]|\right\}.$$

It has been shown in Chan et al. [8] that it is NP-hard to implement any algorithm with *constant* approximation ratio. Although our approximation algorithm has polynominal complexity, it is worth noting that the approximation ratio in Theorem 4 has dependence on the data dimension D and cardinality budget d. It is of research interest to develop approximation algorithms for solving problem (Quad-MMD) with tighter approximation ratio. For instance, Dey et al. [12] have developed a special relaxation formulation for problem (Quad-MMD) with $t=0$ with an optimality gap of $(1+\sqrt{d/(d+1)})^2$, where the relaxation formulation with moderate problem size can be solved in reasonable amount of time. For $t=0$, Chan et al. [8] provide polynomial approximation algorithms different from our SDP formulation that achieve the similar approximation ratio. We leave the extension of their proposed algorithms for general vector t for future study.

REMARK 2 (CONS ABOUT QUADRATIC KERNEL). Under the quadratic kernel choice, it can be shown that

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

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

$$(\mathcal{A}(\mu,\nu))_{k_1,k_2} = \left(\mathbb{E}_{x\sim\mu}[x[k_1]x[k_2]] - \mathbb{E}_{y\sim\nu}[y[k_1]y[k_2]]\right)^2$$

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

$$\mathcal{T}(\mu,\nu)[k] = 2c\left(\mathbb{E}_{x\sim\mu}[x[k]] - \mathbb{E}_{y\sim\nu}[y[k]]\right)^2.$$

In this case, the projected MMD aims to find a direction z such that the difference between the first- and second-order moments of projected target distributions is maximized. When those two moments for data distributions are the same, the quadratic kernel-based MMD does not have enough power for finding representative features to distinguish those two distributions.

4. Algorithms for Gaussian MMD

Finally, we consider the Gaussian kernel choice in (6). Define $Z := zz^{\mathrm{T}} \in \mathbb{S}^+_D$ and $M_{x,y} := \frac{1}{2\gamma}(x-y)(x-y)^{\mathrm{T}} \in \mathbb{S}^+_D$. In this case, the problem (3) can be expressed in terms of $D\times D$ symmetric matrics Z as

$$\begin{aligned} \min_{Z\in\mathbb{S}^+_D}\quad & F(Z) \\ \text{s.t.}\quad & \mathrm{Tr}(Z)=1, \|Z\|_0 \le d^2, \mathrm{rank}(Z)=1, \end{aligned} \tag{Gaussian-MMD}$$

Algorithm 4 Convex-Concave Procedure for solving (13)

Input: Outer maximum iterations T_{out}, Inner maximum iterations T_{in}, step size $\{\gamma_t\}_t$, initial guess Z_1.
for $m = 1, \ldots, T_{\text{out}}$ **do**
  Initialize $Z^{(0)} \leftarrow Z_m$
  for $t = 1, \ldots, T_{\text{in}}$ **do**
    Formulate stochastic subgradient estimator G_t of

$$\hat{F}(Z^{(t-1)}; Z_m) + \lambda \|Z^{(t-1)}\|_1.$$

    Update $Y^{(t)} \leftarrow \exp\left(\log Z^{(t-1)} - \gamma_t G_t\right)$
    Update $Z^{(t)} \leftarrow Y^{(t)} / \|Y^{(t)}\|_{\text{Tr}}$
  end for
  Output $Z_m \leftarrow \frac{1}{T_{\text{in}}} \sum_{t \in [T_{\text{in}}]} Z^{(t)}$.
end for
Return $Z_{T_{\text{out}}}$.

where the objective

$$F(Z) = \frac{2\sum_{i\in[n], j\in[m]} e^{-\langle Z, M_{x_i, y_j}\rangle}}{mn} - \frac{\sum_{i,j\in[n]} e^{-\langle Z, M_{x_i, x_j}\rangle}}{n^2} - \frac{\sum_{i,j\in[m]}^{m} e^{-\langle Z, M_{y_i, y_j}\rangle}}{m^2}. \tag{12}$$

We follow the existing literature [11] to obtain a relaxation of the problem (Gaussian-MMD). We first drop the rank constraint and relax the squared ℓ_0-norm constraint as $\|Z\|_1 \le d$. By selecting λ as the optimal Lagrangian multiplier associated with this constraint, we obtain the penalized reformulation:

$$\begin{aligned} \min_{Z\in\mathcal{Z}_D} \quad & F(Z) + \lambda\|Z\|_1 \\ \text{where} \quad & \mathcal{Z}_D = \{Z \in \mathbb{S}_D^+ : \ \text{Tr}(Z) = 1\}. \end{aligned} \tag{13}$$

Although the objective function F defined in (12) is nonconvex in Z, it can be expressed as the difference between two convex functions, which motivates us to apply the *convex-concave* procedure to solve this formulation. For a given solution Z_0, define the convex approximation of F as

$$\begin{aligned} \hat{F}(Z; Z_0) = \frac{2\sum_{i\in[n], j\in[m]} e^{-\langle Z, M_{x_i, y_j}\rangle}}{mn} - \frac{\sum_{i,j\in[n]} e^{-\langle Z_0, M_{x_i, x_j}\rangle} + \langle \nabla e^{-\langle Z_0, M_{x_i, x_j}\rangle}, Z - Z_0\rangle}{n^2} \\ - \frac{\sum_{i,j\in[m]} e^{-\langle Z_0, M_{y_i, y_j}\rangle} + \langle \nabla e^{-\langle Z_0, M_{y_i, y_j}\rangle}, Z - Z_0\rangle}{m^2}. \end{aligned} \tag{14}$$

Our proposed algorithm for solving (13) consists of outer and inner iterations. At outer iterations, we propose to solve

$$Z_{k+1} = \arg\min_{Z\in\mathcal{Z}} \ \left\{ H_k(Z) \triangleq \hat{F}(Z; Z_k) + \lambda\|Z\|_1 \right\}. \tag{15}$$

At inner iterations, we apply the stochastic mirror descent (SMD) algorithm to obtain near-optimal solutions of (15). We present several notations before outlining the detailed algorithm. Define the von Neumann entropy

$$h(Z) = \sum_{i\in[D]} \lambda_i(Z) \log \lambda_i(Z),$$

where $\lambda_1(Z), \ldots, \lambda_D(Z)$ are the eigenvalues of Z. Next, define the von Neumann Bregman divergence

$$\begin{aligned} V(Z_1, Z_2) &= h(Z_1) - h(Z_2) - \langle Z_1 - Z_2, \nabla h(Z_2)^{\text{T}}\rangle \\ &= \text{Tr}(X \log X - X \log Y). \end{aligned}$$

Define the trace norm

$$\|Z\|_{\mathrm{Tr}} = \sum_{i\in[D]} \lambda_i(Z).$$

For a given feasible solution Z of (15), the stochastic mirror descent algorithm first generates an unbiased gradient estimator, denoted as $G(Z)$, and then update the feasible solution with step size γ as the following:

$$\begin{cases} Y^+ = \exp\left(\log Z - \gamma G(Z)\right), \\ Z^+ = Y^+ / \|Y^+\|_{\mathrm{Tr}}. \end{cases}$$

We summarize the detailed algorithm in Algorithm 4, and provide its convergence guarantees in the following.

Proposition 1 **(Convergence at Inner iterations).** *Assume the subgradient estimator G_t satisfies that*

$$\mathbb{E}\left[\|G_t\|_{op}^2\right] \le M_*^2.$$

Denote $Z_k^* = \arg\min_{Z\in\mathcal{Z}_D} H_k(Z)$. With properly chosen hyper-parameters, the iteration points at inner iterations satisfy

$$\mathbb{E}[H_k(Z^{(t)}) - H_k(Z_k^*)] \le M_* \sqrt{\frac{4V(Z_m, Z_k^*)}{T_{\mathrm{in}}}}.$$

Remark 3 (Convergence at Outer Iterations). Assume the SMD algorithm at inner iterations finds a solution of (15) within negligible optimality gap, then it can be verified that the optimal objective value in (15) is non-increasing with respect to the outer iteration number m. This, together with the fact that the optimal objective value in (15) is lower-bounded, implies the sequence of optimal objective values in (15) converge as m goes to infinity.

Remark 4 (Advantages and Concerns of Gaussian MMD). Since the standard Gaussian kernel is a *characteristic kernel*, it can be shown that $\mathrm{MMD}(\mu, \nu; K_z) = 0$ if and only if $z \circ \mu = z \circ \nu$. In other words, as long as there exists $z \in \mathcal{Z}$ such that $z \circ \mu \neq z \circ \nu$, one can assert that the MMD statistic $\mathrm{MMD}(\mu, \nu; K_z) > 0$. This indicates that the Gaussian kernel is a powerful choice for two-sample variable selection, i.e., it is suitable for arbitrary two distinct distributions. Unfortunately, our proposed algorithm may not succeed in finding a global optimum projection vector of the problem (Gaussian-MMD). It is also of research interest to develop exact algorithms for solving this formulation.

5. Statistical Performance Guarantees

In this section, we provide statistical performance guarantees for the selection coefficient obtained from MMD statistics maximization:

$$\hat{z} = \arg\max_{z\in\mathcal{Z}} S^2(\mathrm{x}^n, \mathrm{y}^m; K_z),$$

though for the Gaussian kernel choice, one may not succeed in finding the corresponding global optimum solution. Throughout this section, we assume $0 \le K_{z^*}(x, y) \le \overline{K}$ for any $x, y \in \Omega$.

Theorem 5 **(Concentration Properties).** *Fix an error probability $\eta \in (0, 1)$, and assume that there exists a projection direction z^* such that $\mathrm{MMD}(\mu, \nu; K_{z^*}) \ge \Delta$. Then the following holds with probability at least $1 - \eta$:*

$$S(\mathrm{x}^n, \mathrm{y}^m; K_{\hat{z}}) \ge \Delta - \epsilon,$$

where the constant

$$\epsilon = 2\left(\left(\frac{\overline{K}}{m}\right)^{1/2} + \left(\frac{\overline{K}}{n}\right)^{1/2}\right) + \sqrt{\frac{2\overline{K}(m+n)}{mn} \log\frac{2}{\eta}}. \tag{16}$$

In comparison with the statistical properties developed in [39, Section 5], our technical assumptions are much milder because we neither assume the sample size $m = n$, distributions are absolutely continuous, nor the sample space is compact. Based on Theorem 5, we provide uncertainty quantification on our testing statistics under null and alternative scenarios.

THEOREM 6 **(PERFORMANCE GUARANTEES AT H_0)**. *Fix a level $\eta \in (0,1)$. Assume that the kernel function satisfies that for any $x, y \in \Omega$ and $z_1, z_2 \in \mathcal{Z}$,*

$$\left|K_{z_1}(x,y) - K_{z_2}(x,y)\right| \le L\|z_1 - z_2\|.$$

Define the sample size $N = n \wedge m$. Under the null hypothesis $H_0 : \mu = \nu$, the following holds with probability at least $1-\eta$:

$$S^2(\mathrm{x}^n, \mathrm{y}^m; K_{\hat{z}}) \lesssim \frac{D}{N}\left[\log\frac{D}{N} + \log\frac{1}{\eta}\right],$$

where $\lesssim$ refers to "less than" with a constant depending only on $(\overline{K}, L)$.

THEOREM 7 **(PERFORMANCE GUARANTEES AT H_1)**. *Fix an error probability $\eta \in (0,1)$. Consider the following assumption for different types of kernel functions:*

- *For linear kernel, assume*
$$\left(\sum_{i\in[d]} (\bar{x}[k_i] - \bar{y}[k_i])^2\right)^{1/4} \ge \Delta,$$
where $k_1, \ldots, k_d$ are indices corresponding to the d-largest elements from $\{(\bar{x}[k] - \bar{y}[k])^2\}_{k\in[D]}$.
- *For quadratic kernel, assume there exists $S \subseteq [D]$ with $|S| = d$ such that either*
$$\lambda_{\max}(\mathcal{A}(\mu,\nu)_{S,S}) \ge \Delta^2$$
or $\|\mathcal{T}(\mu,\nu)_S\|_2 \ge \Delta^2$.
- *For Gaussian kernel, assume there exists a direction $z \in \mathcal{Z}$ such that $\mathrm{MMD}(z\circ\mu, z\circ\nu; K_{Gauss}) \ge \Delta$, where $K_{Gauss}(x,y) = \exp(-\frac{1}{2\gamma}\|x-y\|_2^2)$ is the standard Gaussian kernel.*

Then the relation $S(\mathrm{x}^n, \mathrm{y}^m; \hat{z}) \ge \Delta - \epsilon$ holds with probability at least $1-\eta$, where the constant ϵ is defined in (16).

6. Experiments

We compare the performance of variable selection based on the following approaches: (I) Linear MMD: MMD with a linear kernel, as described in Section 2.1. (II) Quadratic MMD: MMD with a quadratic kernel, as described in Section 3. Here we train the projection vector either using exact or approximation algorithm. (III) Gaussian MMD: MMD with a Gaussian kernel, as described in Section 4. Here we train the projection vector using approximation algorithm. (IV) Sparse Logistic Regression: a framework that trains the projection vector with ℓ_0-norm constraint to minimize the logistic loss [4]. (V) Projected Wasserstein: variable selection framework using projected Wasserstein distance [39]. Since we consider the problem of two-sample testing, we first quantify the performance of variable selection in terms of hypothesis testing metrics rather than the prediction accuracy metrics used in literature (see, e.g., Hazimeh et al. [26], Hastie et al. [25]) in Section 6.1. In Section 6.2, we quantify the performance using *false-discovery proportion* (FDP) and the *non-discovery proportion* (NDP) defined as [2]

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

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. Details about experiment setup together with additional experiments are omitted in Appendix EC.2.

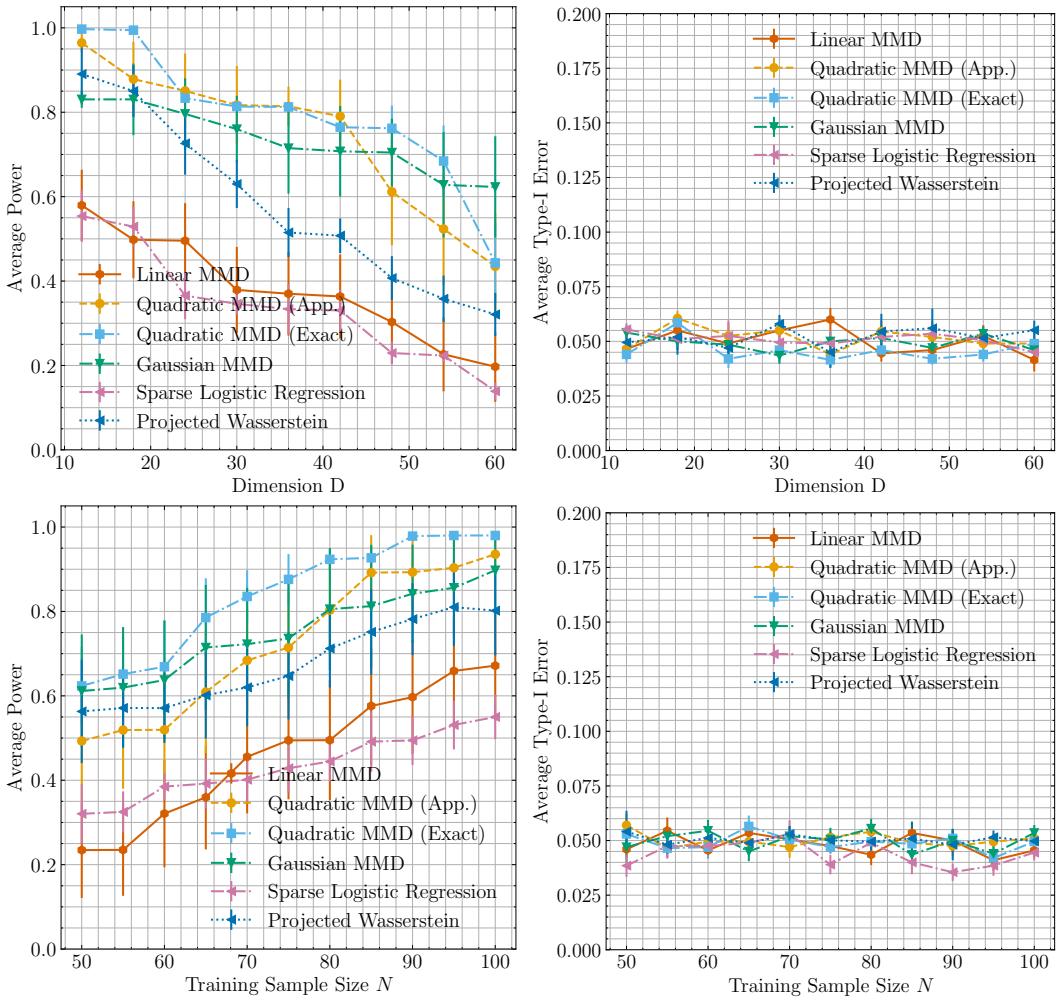


Figure 1 Testing results on Gaussian distributions across different choices of data dimension D and sample size n. Error bars in plots are reproduced with 10 independent trials. For figures from left to right, they correspond to plots for (a) testing power for different D with fixed type-I error $\alpha = 0.05$, (b) average type-I error for different D under H_0, (c) testing power for different n with fixed type-I error $\alpha = 0.05$, and (d) average type-I error for different n under H_0, respectively.

6.1. Study for synthetic dataset

We follow the similar setup in Mueller & Jaakkola [39] to generate the synthetic datasets. Here μ and ν are two multivariate Gaussian distributions of dimension D with different mean vectors and covariance matrices. We take the intrinsic dimension characterizing their differences to be $d = 3$, and the sample size $m = n$. Unless otherwise specified, we take $n = 100$ and $D = 60$. Fig. 1 reports the testing power and type-I error for various approaches across different choices of sample size and data dimension. From the figure, we can see that the logistic regression approach does not have competitive performance. One possible explanation is that the data points from two groups are not linearly separable. Since the linear MMD only utilizes the first-order moment of target distributions to perform testing, we can see it loses power compared with other MMD baselines. Since Gaussian distribution can be fully determined using its first- and second-order moment, one can check the quadratic MMD method with exact algorithm performs the best. Since one cannot solve the Gaussian MMD or projected Wasserstein distance training problem into global optimum, we can see these two methods do not achieve superior performance than quadratic MMD in this example.

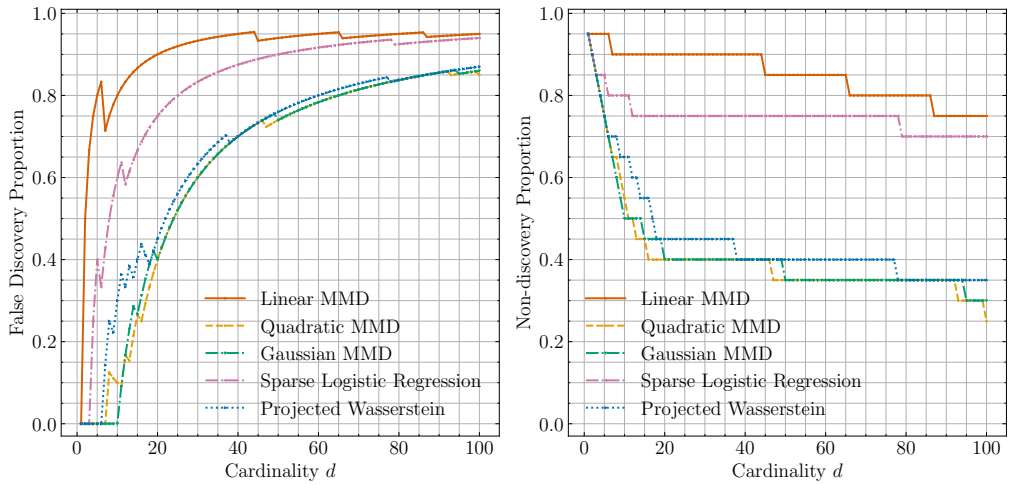


Figure 2 FDP and NDP metrics obtained by various approaches for different choices of cardinality budget d using MADELON dataset.

6.2. Study for MADELON dataset

Next, we examine the performance of variable selection on a large-scale dataset MADELON with sample size $m = n = 1000$ and the data dimension $D = 500$ [23]. Only 20 features within this dataset are useful to detect the differences between two groups. Since the problem size is relatively large, we only use approximation algorithms for quadratic MMD-based variable selection: we select $D' = 200$ features using the greedy algorithm in Algorithm 5 (see Appendix EC.3) as a pre-processing step, and then use the approximation algorithm in Algorithm 3 for variable selection. Fig. 2 reports the FDP and NDP from various approaches with different choices of cardinality constraints. From the plot we can see that quadratic and Gaussian MMD frameworks have lower values of FDP and NDP, which indicate that they outperform other baseline models.

6.3. Study for Healthcare Dataset

Finally, we study the performance of variable selection on a healthcare dataset [50] 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 $m_{\mathrm{Tr}} = 20000, n_{\mathrm{Tr}} = 2000$ and specify the reamining as validation samples. We quantify the performance of variable selection as the testing power on testing samples with sample size $m_{\mathrm{Te}} = n_{\mathrm{Te}} = 100$, which are selected randomly from the validation sample sets.

Fig. 3 reports the top 10 features selected by various approaches based on the training samples. We repeat the experiment of testing for 2000 independent trials and report the averaged testing power for various approaches in Table 1. From the Table, we can see that methods Quadratic MMD (App.) and Quadratic MMD (Exact.) perform the best, and the intersection of those selected features are

pulse , best_map, dbp_cuff, dbp_line, unassisted_resp_rate, end_tidal_co2, hco3, ph.

7. 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

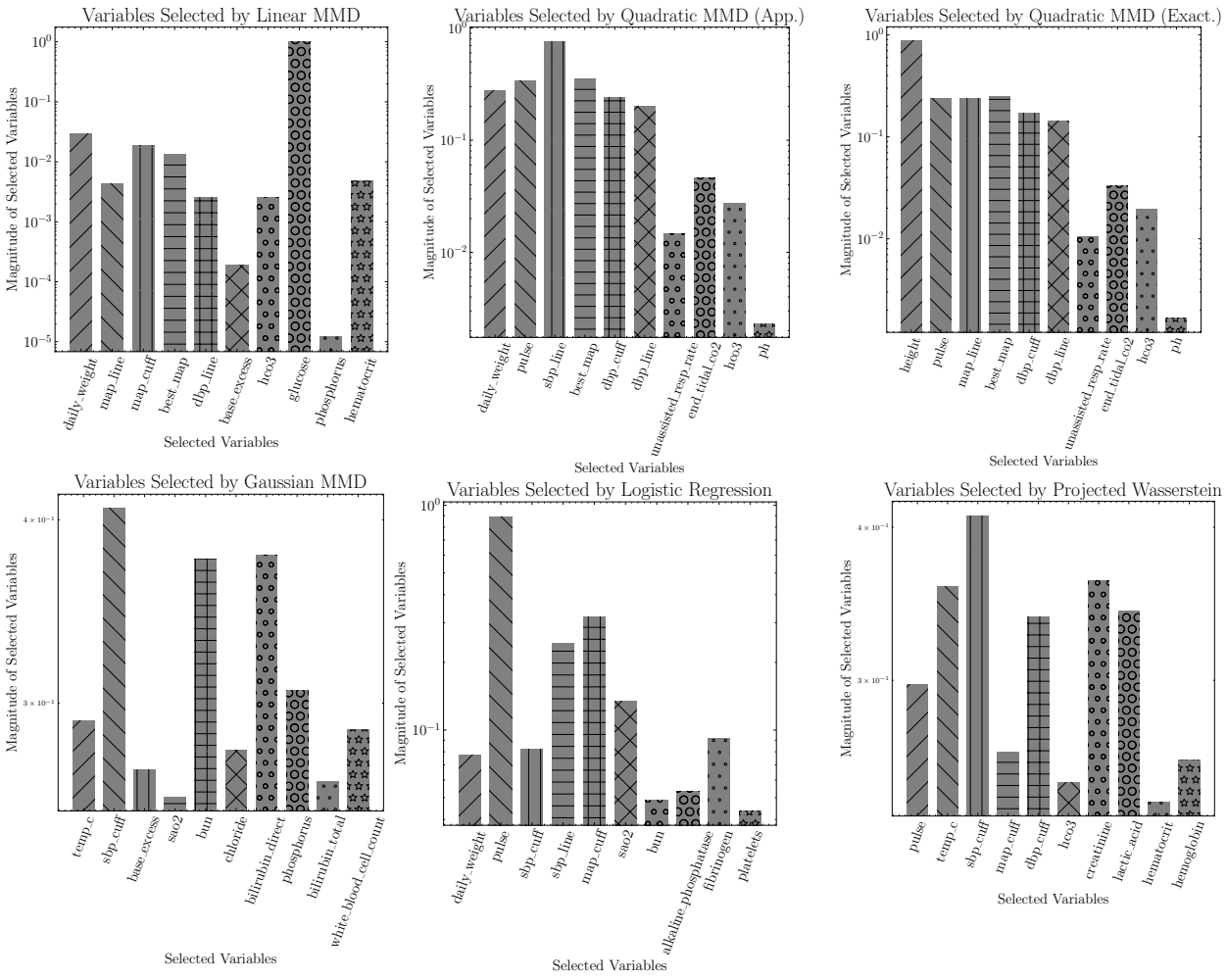


Figure 3 Top 10 variables selected by various approaches in the healthcare dataset

Table 1 Averaged testing power for the healthcare dataset.

Linear MMD	Quadratic MMD (App.)	Quadratic MMD (Exact)	Gaussian MMD	Logistic Regression	Projected Wasserstein
0.435	0.857	0.921	0.794	0.771	0.749

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. First, the kernel bandwidth hyper-parameters have crucial impact on the performance of two-sample testing, while in experiments we specify them using the median heuristic. It is important to study the optimal choice of those hyper-parameters in the variable selection setup. Second, the selection coefficient z is subject to ℓ_2 and ℓ_0 norm constraints in the proposed framework, but extensions to other types of constraints are also possible and bring extra benefits. Finally, it is meaningful to develop more efficient algorithms to solve our proposed formulation.

References

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

[2] Bajwa, W. U. and Mixon, D. G. Group model selection using marginal correlations: The good, the bad and the ugly. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 494–501. IEEE, October 2012.

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

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

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

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

[7] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[8] Chan, S. O., Papailliopoulos, D., and Rubinstein, A. On the approximability of sparse pca. In *Conference on Learning Theory*, pp. 623–646. PMLR, July 2016.

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

[10] Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, December 2015.

[11] d'Aspremont, A., Ghaoui, L., Jordan, M., and Lanckriet, G. A direct formulation for sparse pca using semidefinite programming. *Advances in neural information processing systems*, 17, December 2004.

[12] Dey, S. S., Mazumder, R., and Wang, G. A convex integer programming approach for optimal sparse pca. *arXiv preprint arXiv:1810.09062*, October 2018.

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

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

[15] Dey, S. S., Molinaro, M., and Wang, G. Solving sparse principal component analysis with global support. *Mathematical Programming*, pp. 1–39, July 2022.

[16] Fortin, C. and Wolkowicz, H. The trust region subproblem and semidefinite programming. *Optimization methods and software*, 19(1):41–67, 2004.

[17] Gally, T. and Pfetsch, M. E. Computing restricted isometry constants via mixed-integer semidefinite programming. *preprint, submitted*, 2016.

[18] Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, August 2022.

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

[20] Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

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

[22] Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pp. 1205–1213, December 2012.

[23] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

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

[25] Hastie, T., Tibshirani, R., and Tibshirani, R. Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4):579 – 592, 2020.

[26] Hazimeh, H., Mazumder, R., and Radchenko, P. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *arXiv preprint arXiv:2104.07084*, October 2021.

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

[28] Idé, T., Lozano, A. C., Abe, N., and Liu, Y. Proximity-based anomaly detection using sparse structure learning. In *Proceedings of the 2009 SIAM international conference on data mining*, pp. 97–108. SIAM, 2009.

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

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

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

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

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

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

[35] Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pp. 6316–6326, August 2020.

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

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

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

[39] Mueller, J. and Jaakkola, T. Principal differences analysis: Interpretable characterization of differences between distributions. In *Advances in Neural Information Processing Systems*, volume 28, December 2015.

[40] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, January 2009.

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

[42] Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, August 2021.

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

[44] Sun, Z. and Zou, S. Kernel robust hypothesis testing. *arXiv preprint arXiv:2203.12777*, March 2022.

[45] Taguchi, G. and Rajesh, J. New trends in multivariate diagnosis. *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 233–248, August 2000.

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

[47] Wang, J., Chen, M., Zhao, T., Liao, W., and Xie, Y. A manifold two-sample test study: Integral probability metric with neural networks. *arXiv preprint arXiv:2205.02043*, May 2022.

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

[49] Wang, J., Gao, R., and Xie, Y. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, December 2022.

[50] Wang, J., Moore, R., Xie, Y., and Kamaleswaran, R. Improving sepsis prediction model generalization with optimal transport. In *Machine Learning for Health*, pp. 474–488. PMLR, November 2022.

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

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

EC.1. Proofs of Technical Results

EC.1.1. Proof in Section 2.1

Proof of Theorem 1 For each $k \in [D]$, we define the binary variable $q[k] = 1$ if the k-th feature is selected, and $q[k] = 0$ otherwise. By linearizing the zero-norm constraint using the binary vector q, we can reformulate (Linear-MMD) as

$$\max_{z\in\mathbb{R}^D, q\in\mathcal{Q}} \left\{ a^{\mathrm{T}} z : \ \|z\|_2 \le 1, \left|z[k]\right| \le q[k], k \in [D] \right\}, \tag{EC.1}$$

where the set

$$\mathcal{Q} = \left\{ q \in \mathbb{F}^D : \quad \sum_{k\in[D]} q[k] = d \right\}.$$

Given a size-d set $S \subseteq [D]$, one can check

$$\max_{z\in\mathbb{R}^D} \left\{ a^{\mathrm{T}} z : \ \|z\|_2 \le 1, z[k] = 0, \forall k \notin S \right\} = \max_{z\in\mathbb{R}^d} \left\{ a_S^{\mathrm{T}} z : \ \|z\|_2 \le 1 \right\} = \|a_S\|_2,$$

where the associated optimal solution z^* satisfies

$$z^*[k] = \begin{cases} a[k]/\|a_S\|_2, & \text{if } k \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{EC.2}$$

Hence, we obtain a combinatorial reformulation of (EC.1):

$$\max_{S} \left\{ \|a_S\|_2 : \quad |S| \le d, S \subseteq [D] \right\}. \tag{EC.3}$$

Let $\{k_1, \ldots, k_D\}$ be the sorted indices of $\{1, \ldots, D\}$ such that $a[k_1] \ge \cdots \ge a[k_D]$. Also, it should be noted that $a[k] \ge 0$ for $k \in [D]$. Combining these two facts, we can assert that an optimal solution to the formulation (EC.3) is

$$S^* = \{k_1, \ldots, k_d\}. \tag{EC.4}$$

By (EC.2) and (EC.4), we obtain the closed-form expression on the optimal solution to (Linear-MMD).

EC.1.2. Proofs in Section 3

Proof of Theorem 2 (I) A natural combinatorial reformulation of (Quad-MMD) is

$$\max_{\substack{S\subseteq[D]:\ |S|\le d\\ z\in\mathbb{R}^D}} \left\{z^{\mathrm{T}}Az + z^{\mathrm{T}}t:\ \|z\|_2 = 1, z[k] = 0, \forall k\notin S\right\}. \tag{EC.5}$$

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

$$\begin{aligned}&\max_{z\in\mathbb{R}^D} \left\{z^{\mathrm{T}}Az + z^{\mathrm{T}}t:\ \|z\|_2 = 1, z[k] = 0, \forall k\notin S\right\}\\ =&\max_{z\in\mathbb{R}^d} \left\{z^{\mathrm{T}}A_{S,S}z + z^{\mathrm{T}}t_S:\ \|z\|_2 = 1\right\}.\end{aligned} \tag{EC.6}$$

Next, we linearize the problem (EC.6) 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.6) as

$$\begin{aligned}\max_{Z\in\mathbb{S}^+_{d+1}}\quad & \langle\tilde{A}_{S,S}, Z\rangle\\ \text{s.t.}\quad & \mathrm{rank}(Z) = 1,\\ & Z_{0,0} = 1, \mathrm{Tr}(Z) = 2.\end{aligned} \tag{EC.7}$$

In particular, constraints $Z\succeq 0, \mathrm{rank}(Z) = 1, Z_{0,0} = 1$ imply $Z = \begin{pmatrix}1 & z^{\mathrm{T}}\\ z & zz^{\mathrm{T}}\end{pmatrix}$ for some vector $z\in\mathbb{R}^d$, and the condition $\mathrm{Tr}(Z) = 2$ implies $\|z\|_2 = 1$. By [33, Corollary 3], we further obtain the following equivalent reformulation of problem (EC.6) when dropping the nonconvex rank constraint $\mathrm{rank}(Z) = 1$:

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

In summary, we obtain the following reformulation of (Quad-MMD):

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

It remains to show the equivalence between formulations (7) and (EC.9). We only need to show for any feasible $q\in\mathcal{Q}$ with its support $S := \{k:\ q[k] = 1\}$, it holds that

$$\max_{Z\in\mathbb{S}^+_{D+1}} \left\{\langle\tilde{A}, Z\rangle:\ Z_{i,i}\le q[i], i\in[D], Z_{0,0} = 1, \mathrm{Tr}(Z) = 2\right\} = \max_{Z\in\mathbb{S}^+_{d+1}} \left\{\langle\tilde{A}_{S,S}, Z\rangle:\ Z_{0,0} = 1, \mathrm{Tr}(Z) = 2\right\}. \tag{EC.10}$$

Since $Z\in\mathbb{S}^+_{D+1}$ is a postive 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.10) indeed holds true.

(II) The continuous relaxation of problem (7) becomes

$$\max_{Z\in\mathbb{S}^+_{D+1},\, q\in\overline{\mathcal{Q}}} \left\{\langle \tilde{A}, Z\rangle :\ Z_{i,i}\le q[i], i\in[D], Z_{0,0}=1, \mathrm{Tr}(Z)=2\right\}.$$

Since $Z_{0,0}=1, \mathrm{Tr}(Z)=2$, the linking constraint $Z_{i,i}\le q[i]$ is *redundant* for $i\in[D]$. As a result, the problem above reduces to

$$\max_{Z\in\mathbb{S}^+_{D+1}} \left\{\langle \tilde{A}, Z\rangle :\ Z_{0,0}=1, \mathrm{Tr}(Z)=2\right\}.$$

According to Fortin & Wolkowicz [16], it corresponds to the exact SDP reformulation of the trust region subproblem in (8).

Proof of Corollary 1 It remains to show additional constraints (9a), (9b) are indeed valid for (7).

- We highlight that $Z_{1:D,1:D}\succeq 0$ and $\mathrm{Tr}(Z_{1:D,1:D})=1$, i.e., the technical assumptions in [17, Lemma 1] hold. By [17, Lemma 1], one can assert that (9a) holds.
- From the proof of Theorem 2, one can assume $Z=\begin{pmatrix}1 & z^{\mathrm{T}}\\ z & zz^{\mathrm{T}}\end{pmatrix}$ without loss of generality, where z is also feasible to (Quad-MMD). It is known that $\|z\|_2\le\sqrt{d}$ (see, e.g., [14]) and therefore

$$\sum_{j\in[D]}|Z_{i,j}| = \sum_{j\in[D]}|z_i||z_j| \le \sqrt{d}|z_i| \le \sqrt{d},$$

where the last inequality is because $Z_{i,i}=z_i^2\le 1, i\in[D]$. Hence, the relation (9b) also holds.

Proof of Theorem 3 We re-write $f(q)$ as the following optimization problem:

$$\begin{aligned} f(q) = \max_{Z\in\mathbb{S}^+_{D+1},\, U}\quad & \langle \tilde{A}, Z\rangle & \\ & Z_{0,0}=1, & [\lambda_0]\\ & \mathrm{Tr}(Z)=2, & [\lambda]\\ & -U_{i,j}\le Z_{i,j}\le U_{i,j}, \forall i,j\in[D], & [\alpha^-_{i,j},\alpha^+_{i,j}]\\ & U_{i,j}\le M_{i,j}q[i], \forall i,j\in[D], & [\sigma_{i,j}]\\ & \sum_{j\in[D]} U_{i,j}\le\sqrt{d}q[i], \forall i\in[D], & [\beta[i]] \end{aligned}$$

Here we associate dual variables with primal constraints in brackets. Its Lagrangian dual reformulation becomes

$$\begin{aligned} \min_{\substack{\lambda,\lambda_0\\ \alpha^-,\alpha^+,\beta,\sigma\ge 0}}\ \max_{Z\in\mathbb{S}^+_{D+1},\, U}\quad & \langle \tilde{A}, Z\rangle + \lambda_0\left(1-Z_{0,0}\right) + \lambda_1\left(2-\mathrm{Tr}(Z)\right)\\ & + \sum_{i,j}\alpha^+_{i,j}\left[U_{i,j}-Z_{i,j}\right] + \sum_{i,j}\alpha^-_{i,j}\left[U_{i,j}+Z_{i,j}\right] + \sum_{i,j}\sigma_{i,j}\left[M_{i,j}q[i]-U_{i,j}\right] + \sum_i\beta_i\left[\sqrt{d}q[i]-\sum_{j\in[D]}U_{i,j}\right]. \end{aligned}$$

Or equivalently, it can be written as

$$\begin{aligned} \min_{\substack{\lambda,\lambda_0\\ \alpha^-,\alpha^+,\beta,\sigma\ge 0}}\quad & \lambda_0+2\lambda_1+\sum_{i,j}\sigma_{i,j}M_{i,j}q[i]+\sum_i\beta[i]\sqrt{d}q[i] + \max_{Z\in\mathbb{S}^+_{D+1},\, U}\quad \langle \tilde{A}, Z\rangle - \lambda_0 Z_{0,0} - \lambda_1\mathrm{Tr}(Z)\\ & + \sum_{i,j}\alpha^+_{i,j}\left[U_{i,j}-Z_{i,j}\right] + \sum_{i,j}\alpha^-_{i,j}\left[U_{i,j}+Z_{i,j}\right] - \sum_{i,j}\sigma_{i,j}U_{i,j} - \sum_{i,j}\beta_i U_{i,j}. \end{aligned}$$

Solving the inner maximization subproblem with respect to Z and U yields the dual problem

$$\begin{aligned}\min_{\substack{\lambda,\lambda_0\\ \alpha^-,\alpha^+,\beta,\sigma\geq 0}} \quad & \lambda_0+2\lambda_1+\sum_{i,j}\sigma_{i,j}M_{i,j}q[i]+\sum_i \beta_i\sqrt{d}q[i]\\ \text{s.t.} \quad & \begin{pmatrix}\lambda_0 & 0\\ 0 & \lambda_1 I_D+\alpha^+-\alpha^-\end{pmatrix}\succeq \tilde{A}\\ & \sigma_{i,j}+\beta[i]\geq \alpha^+_{i,j}+\alpha^-_{i,j},\quad \forall i,j\in[D].\end{aligned}$$

We eliminate σ from the dual problem by optimizing $\sigma_{i,j}$ and set

$$\sigma_{i,j}=\max\left(0,\alpha^+_{i,j}+\alpha^-_{i,j}-\beta[i]\right).$$

Proof of Theorem 4 The left-hand-side relation is easy to show. The proof for the right-hand-side relation is separated into two parts: (I) $\mathsf{optval}(11)\leq \|t\|_2+d\cdot\{\mathsf{optval}(7)-\min_k|t[k]|\}$; (II) $\mathsf{optval}(11)\leq \|t\|_2+D/d\cdot\mathsf{optval}(7)$.

(I) First, we construct feasible solutions Z in (7):

$$Z_{0,0}=1, Z_{k,k}=1, Z_{0,k}=Z_{k,0}=\pm 1,$$

and otherwise $Z_{i,j}=0$. Based on this observation, it holds that

$$\mathsf{optval}(7)\geq A_{k,k}+\left|t[k]\right|,\quad \forall k\in[D].$$

Or equivalently,

$$A_{k,k}\leq \mathsf{optval}(7)-\min_k\left|t[k]\right|.$$

For any feasible solution (Z,q) in (11), we define the matrix and vector

$$Z'=Z_{1:D,1:D},\quad z=Z_{0,1:D}.$$

Then the objective in (11) can be written as

$$\langle Z,\tilde{A}\rangle=z^{\mathrm{T}}t+\langle Z',A\rangle.$$

On the one hand,

$$z^{\mathrm{T}}t\leq\sum_{i\in[D]}|t[i]|\sqrt{Z_{i,i}}\leq\sqrt{\sum_{i\in[D]}t[i]^2}\sqrt{\sum_{i\in[D]}Z_{i,i}}=\|t\|_2,$$

where the first inequality is due to taking absolute values and the fact that $|Z_{0,i}|\leq\sqrt{Z_{0,0}Z_{i,i}}=\sqrt{Z_{i,i}}$, the second inequality is based on the Cauchy-Schwarz inequality, and the last equality is based on the condition $\mathrm{Tr}(Z)=2$. On the other hand, it holds that

$$\begin{aligned}\langle Z',A\rangle=\sum_{i,j}Z'_{i,j}A_{i,j}&\leq\max_{i,j}\ |A_{i,j}|\cdot\sum_{i,j}|Z_{i,j}|\\ &\leq\sum_{i,j}|Z'_{i,j}|\cdot\left\{\mathsf{optval}(7)-\min_k|t[k]|\right\}\\ &\leq d\cdot\left\{\mathsf{optval}(7)-\min_k|t[k]|\right\}.\end{aligned}$$

where the first inequality is due to taking absolute values, the second is due to the fact that

$$|A_{i,j}|\leq\sqrt{A_{i,i}A_{j,j}}\leq\max_{k\in[D]}\ A_{k,k}\leq\mathsf{optval}(7)-\min_k\left|t[k]\right|,$$

and the last is due to the fact that

$$\sum_{i,j} |Z'_{i,j}| \le \sum_i \sqrt{d}\sqrt{Z'_{i,i} q[i]} \le \sqrt{d}\sqrt{\sum_i Z'_{i,i}}\sqrt{\sum_i q[i]} = d.$$

Combining those two relations, we obtain the relation

$$\langle Z, \tilde{A}\rangle \le \|t\|_2 + d \cdot \left\{ \mathsf{optval}(7) - \min_k |t[k]| \right\}, \qquad \forall (Z,q).$$

Therefore, we obtain

$$\mathsf{optval}(11) \le \|t\|_2 + d \cdot \left\{ \mathsf{optval}(7) - \min_k |t[k]| \right\}.$$

(II) For any feasible solution (Z,q) in (7), we enforce $Z_{0,1:D} = Z_{1:D,0} = 0$, then the updated solution is still feasible, with the associated objective value

$$\langle Z_{1:D,1:D}, A\rangle.$$

Therefore, we obtain the relation

$$\mathsf{optval}(7) \ge \max_{Z\in\mathbb{S}^+_D, q\in\mathcal{Q}} \left\{ \langle Z, A\rangle : \; Z_{i,i} \le q[i], i \in [D], \mathrm{Tr}(Z) = 1 \right\} \ge d/D \cdot \lambda_{\max}(A),$$

where the last inequality is due to [31, Proposition 2 and proof of Theorem 5]. For any feasible solution (Z,q) in (11), we define the matrix Z' and vector z similar as in Part (I). On the one hand, $z^{\mathrm{T}} t \le \|t\|_2$ still holds. On the other hand,

$$\langle Z', A\rangle \le \max_{Z'\in\mathbb{S}^+_D} \left\{ \langle Z', A\rangle : \; \mathrm{Tr}(Z') = 1 \right\} = \lambda_{\max}(A) \le D/d \cdot \mathsf{optval}(7)$$

Therefore, we obtain

$$\mathsf{optval}(11) \le \|t\|_2 + D/d \cdot \mathsf{optval}(7).$$

EC.1.3. Proof in Section 4

Consider the minimization of the objective function $F(\theta) = \mathbb{E}[f_\theta(z)]$ with $\theta \in \Theta$. In particular, we assume the constraint set Θ is non-empty, closed and convex. We also impose the following assumption regarding the (sub-)gradient oracles when using the SMD algorithm:

Assumption EC.1 **(Stochastic Oracles of Gradient Estimate)**. *The objective function $F(\theta)$ is convex in θ, and we have the stochastic oracle such that for given θ we can generate a stochastic vector $G(\theta, \xi)$ such that $\mathbb{E}[G(\theta, \xi)] \in \partial F(\theta)$, where $\partial F(\theta)$ is the subdifferential set of $F(\cdot)$ at θ. Also, suppose there exists a constant $M_* > 0$ so that*

$$\mathbb{E}\left[\|G(\theta, \xi)\|_*^2\right] \le M_*^2, \quad \forall \theta \in \Theta.$$

Under the above assumption, the SMD algorithm generates the following iteration:

$$\theta_{t+1} = \mathrm{Prox}_{\theta_t}\left(\gamma_t G(\theta_t, \xi^t)\right), \quad \theta_1 \in \Theta, \quad t = 1, \ldots, T-1.$$

Here the operator $\mathrm{Prox}_\theta(\cdot)$ denotes the prox-mapping induced from a distance generating function that is κ-strongly convex. For simplicity of discussion, we employ constant step size policy $\gamma_t := \gamma$ for $t = 1, \ldots, T-1$. Similar results can be found in Shapiro et al. [42], Nemirovski et al. [40], Wang et al. [49].

Lemma EC.1 **(SMD for Nonsmooth Convex Optimization)**. *Under Assumption EC.1, let the estimation of optimal solution at the iteration j be*

$$\tilde{\theta}_{1:j} = \frac{1}{j}\sum_{t=1}^{j} \theta_t.$$

When taking constant step size

$$\gamma = \sqrt{\frac{2\kappa V(\theta_1, \theta^*)}{T M_*^2}},$$

it holds that

$$\mathbb{E}\left[F(\tilde{\theta}_{1:T}) - F(\theta^*)\right] \le M_* \sqrt{\frac{2V(\theta_1, \theta^*)}{\kappa T}}.$$

Proof of Proposition 1 This proposition follows directly from Lemma EC.1 and the fact that the von Neumann entropy is 1/2-strongly convex.

EC.1.4. Proofs in Section 5

To show the proofs in Section 5, we rely on the following technical lemma on concentration properties of standard MMD.

LEMMA EC.2 **(MEASURE CONCENTRATION OF MMD [21, THEOREM 7])**. *Let p, q be two probability distributions, and $\hat{p}_m, \hat{q}_n$ be the empirical measures from m and n samples generated from p, q, respectively. Suppose the MMD function is constructed based on the kernel function $k(x, y)$ with $0 \le k(x, y) \le K$. Then it holds that*

$$\Pr\left\{\left|MMD(p,q) - MMD(\hat{p}_m, \hat{q}_n)\right| > 2(K/n)^{1/2} + 2(K/m)^{1/2} + \epsilon\right\} \le 2\exp\left(\frac{-\epsilon^2 mn}{2K(m+n)}\right).$$

Proof of Theorem 5 Fix error

$$\delta = 2\left(\left(\frac{K}{m}\right)^{1/2} + \left(\frac{K}{n}\right)^{1/2}\right) + \sqrt{\frac{2K(m+n)}{mn}\log\frac{2}{\eta}}.$$

Since the projection vector $\hat{z}$ maximizes $S^2(\mathrm{x}^n, \mathrm{y}^m; K_z)$ and therefore maximizes $S(\mathrm{x}^n, \mathrm{y}^m; K_z)$ over $z \in \mathcal{Z}$, it holds that

$$\begin{aligned}
&\Pr\left\{S(\mathrm{x}^n, \mathrm{y}^m; K_{\hat{z}}) \ge \Delta - \delta\right\} \\
\ge &\Pr\left\{S(\mathrm{x}^n, \mathrm{y}^m; K_{z^*}) \ge \Delta - \delta\right\} \\
= &\Pr\left\{\Delta - S(\mathrm{x}^n, \mathrm{y}^m; K_{z^*}) \le \delta\right\}.
\end{aligned}$$

Based on our assumption, the event

$$\left\{\Delta - S(\mathrm{x}^n, \mathrm{y}^m; K_{z^*}) \le \delta\right\} \supseteq \left\{\mathrm{MMD}(\mu, \nu; K_{z^*}) - S(\mathrm{x}^n, \mathrm{y}^m; K_{z^*}) \le \delta\right\}.$$

Therefore

$$\begin{aligned}
&\Pr\left\{S(\mathrm{x}^n, \mathrm{y}^m; K_{\hat{z}}) \ge \Delta - \delta\right\} \\
\ge &\Pr\left\{\mathrm{MMD}(\mu, \nu; K_{z^*}) - S(\mathrm{x}^n, \mathrm{y}^m; K_{z^*}) \le \delta\right\} \\
\ge &1 - \eta,
\end{aligned}$$

where the last inequality follows from Lemma EC.2.

The proof of Theorem 6 relies on the following technical lemma.

LEMMA EC.3 **(LIPSCHITZ CONTINUITY ON EMPIRICAL MMD STATISTIC)**. *Assume that the kernel function satisfies that for any $x, y \in \Omega$ and $z_1, z_2 \in \mathcal{Z}$,*

$$\left|K_{z_1}(x, y) - K_{z_2}(x, y)\right| \le L\|z_1 - z_2\|_2.$$

Then

$$\left|S^2(\mathrm{x}^n, \mathrm{y}^m; K_{z_1}) - S^2(\mathrm{x}^n, \mathrm{y}^m; K_{z_2})\right| \le 4L\|z_1 - z_2\|_2.$$

Proof of Lemma EC.3 By direct calculation, it holds that

$$\begin{aligned}
&|S^2(\mathrm{x}^n, \mathrm{y}^m; K_{z_1}) - S^2(\mathrm{x}^n, \mathrm{y}^m; K_{z_2})| \\
\le &\frac{1}{n^2}\sum_{i,j\in[n]}\left|K_{z_1}(x_i, x_j) - K_{z_2}(x_i, x_j)\right| + \frac{1}{m^2}\sum_{i,j\in[m]}\left|K_{z_1}(y_i, y_j) - K_{z_2}(y_i, y_j)\right| \\
&\quad + \frac{1}{mn}\sum_{i\in[n], j\in[m]}\left|K_{z_1}(x_i, y_j) - K_{z_2}(x_i, y_j)\right| \\
&\le 4L\|z_1 - z_2\|_2.
\end{aligned}$$

Proof of Theorem 6 We take a fine grid of points $\{z_1, \ldots, z_N\}$ such that it forms an ρ-net cover of $\mathcal{Z}$. It can be shown that the cardinality

$$N \le \left(1 + \frac{2}{\rho}\right)^D$$

and the population statistic $\mathrm{MMD}(\mu, \nu; K_{z_i}) = 0$ for $i \in [N]$. By the concentration property of the MMD statistics, with probability at least $1 - \eta$, it holds that

$$S(\mathrm{x}^n, \mathrm{y}^m; K_{z_i}) < \epsilon.$$

Based on the union bound, with probability at least $1 - N\eta$, it holds that

$$S(\mathrm{x}^n, \mathrm{y}^m; K_{z_i}) < \epsilon, \quad \forall i \in [N].$$

Equivalently, with probability at least $1 - \eta$, it holds that

$$S(\mathrm{x}^n, \mathrm{y}^m; K_{z_i}) < 2\left(\left(\frac{\overline{K}}{m}\right)^{1/2} + \left(\frac{\overline{K}}{n}\right)^{1/2}\right) + \sqrt{\frac{2\overline{K}(m+n)}{mn} \log\left[\frac{2}{\eta}\left(1 + \frac{2}{\rho}\right)^D\right]}.$$

By definition, there exists a projection vector z from $\{z_1, \ldots, z_S\}$ such that $\|z - \hat{z}\|_2 < \rho$. Then by Lemma EC.3, with probability at least $1 - \eta$, it holds that

$$\begin{aligned}
S^2(\mathrm{x}^n, \mathrm{y}^m; K_{\hat{z}}) &\le \left\{2\left(\left(\frac{\overline{K}}{m}\right)^{1/2} + \left(\frac{\overline{K}}{n}\right)^{1/2}\right) + \sqrt{\frac{2\overline{K}(m+n)}{mn} \log\left[\frac{2}{\eta}\left(1 + \frac{2}{\rho}\right)^D\right]}.\right\}^2 + 4L\rho \\
&\le 2\left\{2\left(\left(\frac{\overline{K}}{m}\right)^{1/2} + \left(\frac{\overline{K}}{n}\right)^{1/2}\right)\right\}^2 + \frac{4\overline{K}(m+n)}{mn} \log\left[\frac{2}{\eta}\left(1 + \frac{2}{\rho}\right)^D\right] + 4L\rho \\
&\le \frac{16\overline{K}(m+n)}{mn} + \frac{4\overline{K}(m+n)}{mn} \log\frac{2}{\eta} + \frac{4\overline{K}D(m+n)}{mn} \log\left(1 + \frac{2}{\rho}\right) + 4L\rho.
\end{aligned}$$

The key is to select the optimal radius ρ to minimize the right-hand-side above. We take $\rho = \frac{\overline{K}D(m+n)}{Lmn}$ and therefore

$$S^2(\mathrm{x}^n, \mathrm{y}^m; K_{\hat{z}}) \lesssim \frac{D}{N} \log\frac{D}{N} + \frac{D}{N} \log\frac{1}{\eta},$$

where $\lesssim$ refers to "less than" with a constant depending only on $(\overline{K}, L)$.

Proof of Theorem 7 By Theorem 5, it remains to show there exists a projection direction z^* such that $\mathrm{MMD}(\mu, \nu; K_{z^*}) \ge \Delta$ for various choices of kernel functions.

- For linear kernel case, we find

$$\mathrm{MMD}(\mu, \nu; K_z) = \left[\sum_{k \in [D]} z[k]\left(\overline{x}[k] - \overline{y}[k]\right)^2\right]^{1/2}.$$

  We take the projection direction z^* such that

$$z^*[k_i] = \frac{\left(\overline{x}[k_i] - \overline{y}[k_i]\right)^2}{\sum_{i' \in [d]}\left(\overline{x}[k_{i'}] - \overline{y}[k_i]\right)^2}, \quad i \in [d],$$

  and $z^*[i] = 0$ for $i \notin \{k_1, \ldots, k_d\}$.

- For quadratic kernel case, we find

$$\mathrm{MMD}(\mu, \nu; K_z) = \left(z^{\mathrm{T}} \mathcal{A}(\mu, \nu) z + z^{\mathrm{T}} \mathcal{T}(\mu, \nu) \right)^{1/2}.$$

  Suppose there exists $S \subseteq [D]$ with $|S| = d$ such that $\lambda_{\max}(\mathcal{A}(\mu, \nu)_{S,S}) \geq \Delta^2$, then we construct z^* so that z^*_S is the leftmost eigenvector of $\mathcal{A}(\mu, \nu)_{S,S}$ satisfying $(z^*)^{\mathrm{T}} \mathcal{T}(\mu, \nu) > 0$. On the other case, we construct z^* such that $z^*_S = \frac{\mathcal{T}(\mu,\nu)_S}{\|\mathcal{T}(\mu,\nu)_S\|_2}$. Both cases lead to

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

- For Gaussian kernel case, we find

$$\mathrm{MMD}(\mu, \nu; K_z) = \mathrm{MMD}(z \circ \mu, z \circ \nu; K_{\mathrm{Gauss}}) \geq \Delta,$$

  and the assumption directly follows.

The proof is completed.

EC.2. Experiment Details

EC.2.1. Details on Synthetic Datasets Generation

We specify the features of the distributions μ and ν to be multivariate Gaussian distributions in blocks of 3, where within each block, common mean parameters are sampled from the uniform distribution on unit sphere and common covariance parameters are sampled from the Wishart distribution with degrees of freedom 3. Only for the first block of 3 features, do we sample separate mean-covariance parameters for μ and separate mean-covariance parameters for ν. Hence, all differences between those two distributions lie in the first $d = 3$ features. To generate data samples in dimension $D = \ell \times d$, we concatenate ℓ of our blocks together, but we always keep the distributions from two groups within the first block different.

EC.2.2. Configurations of Baseline Models

We use the algorithm outlied in Mueller & Jaakkola [39] to implement the projected Wasserstein baseline approach. In the following, we specify the details of logistic regression-based baseline in our experiments.

Training Phase: We optimize the logistic regression formulation:

$$\min_{w\in\mathbb{R}^D, b\in\mathbb{R}} \left\{ \sum_{i=1}^{n_{\mathrm{Tr}}} \ell(y_i, w^{\mathrm{T}} x_i + b) + \frac{1}{2\gamma}\|w\|_2^2 :\ \|w\|_0 \le d \right\},$$

where the feature set $\{x_i\}$ is the concatenation of training sample points from two groups and the label set $\{y_i\}$ corresponds to labels generated either from μ or ν. We follow the outer approximation technique in Bertsimas et al. [4] to solve this formulation. Specially, it is equivalent to

$$\min_{q\in\overline{\mathcal{Q}}} \ c(q),$$

where $c(q) := \max_{\alpha\in\mathbb{R}^n:\ e^{\mathrm{T}}\alpha=0} \ f(\alpha, q)$ with

$$f(\alpha, q) = -\sum_{i\in[n_{\mathrm{Tr}}]} \hat{\ell}(y_i, \alpha_i) - \frac{\gamma}{2} \sum_{j\in[n_{\mathrm{Tr}}]} s_j \alpha^{\mathrm{T}} X_j X_j^{\mathrm{T}} \alpha.$$

Iteratively, we update

$$q^{(t+1)}, \eta^{(t+1)} \leftarrow \arg\min_{s,\eta} \ \left\{ \eta :\ q \in \overline{\mathcal{Q}}, \eta \ge c(q^{(i)}) + \nabla c(q^{(i)})^{\mathrm{T}} (q - q^{(i)}), i \in [t] \right\}.$$

Testing Phase: After the optimal classifer (w^*, b^*) is obtained, we formulate the testing statistic

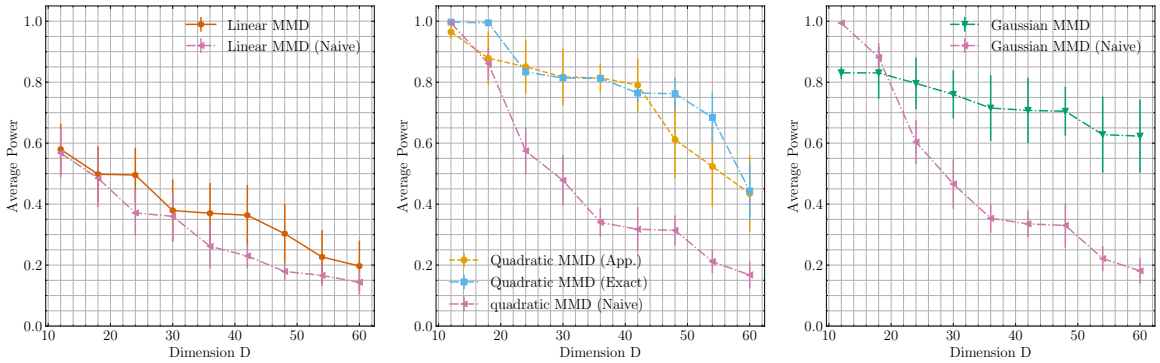
$$\hat{T} = \frac{1}{|X_{\mathrm{Te}}|} \sum_{x\in X_{\mathrm{Te}}} ((w^*)^{\mathrm{T}} x + b^*) - \frac{1}{|Y_{\mathrm{Te}}|} \sum_{x\in Y_{\mathrm{Te}}} ((w^*)^{\mathrm{T}} x + b^*).$$

We reject H_0 if $\hat{T}$ is larger than a certain threshold, where the threshold can be computed using the permutation test similar to that in Algorithm 1. Such a methodology in the testing phase is originally proposed in Cheng & Cloninger [9].

EC.2.3. Numerical Study on Kernel Testing with/without Variable Selection

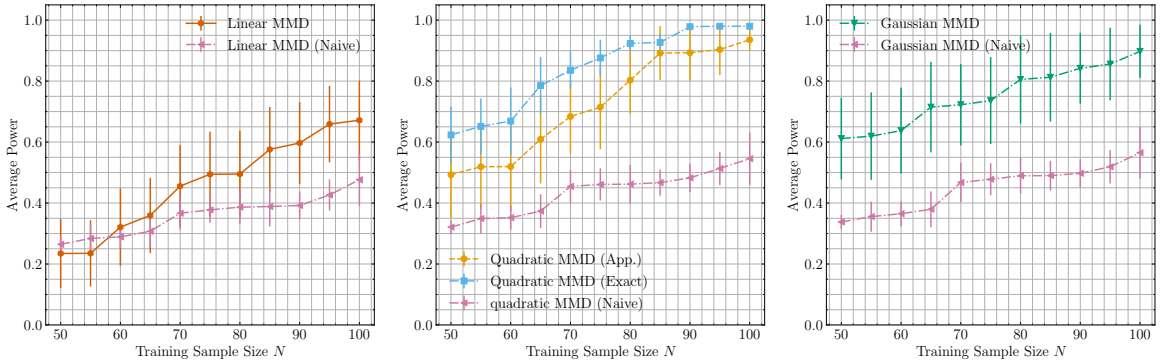
In this subsection, we study the performance of MMD two-sample tests *with variable selection* compared to those tests *without variable selection*. For baseline tests without variable selection, we call it xxxxxx MMD (Naive) with xxxxx $\in$ {Linear, Quadratic, Gaussian}. To implement those baseline methods, we run the two-sample tests using both training and testing data points and specify kernel bandwidth hyper-parameters using *median heuristic*, i.e., they are specified as the median of the differences between two groups of data points. We examine the performance using the same synthetic datasets as in Section 6.1.

Fig. EC.1 and EC.2 report testing results across different choices of data dimension D and training sample size N, respectively. Within each figure, those three subplots correspond to linear, quadratic, and Gaussian kernel functions, respectively. From the plots we realize that for almost all problem instances, the MMD tests with variable selection can improve the performance of the naive MMD tests.



(a) Linear MMD (b) Quadratic MMD (c) Gaussian MMD

Figure EC.1 Testing results for MMD framework with variable selection versus that without variable selection. Here we fix the sample size and vary the data dimension.



(a) Linear MMD (b) Quadratic MMD (c) Gaussian MMD

Figure EC.2 Testing results for MMD framework with variable selection versus that without variable selection. Here we fix the data dimension and vary the sample size.

EC.3. Additional Approximation Algorithms for Solving (Quad-MMD)

Recall the combinatorial reformulation of (Quad-MMD) is

$$\max_{S\subseteq[D]:\ |S|\le d} \Lambda(S),$$

where the set function Λ is defined as

$$\Lambda(S) = \max_{z\in\mathbb{R}^D} \left\{ z^{\mathrm{T}} A z + z^{\mathrm{T}} t :\ \|z\|_2 = 1, \quad z[i] = 0, \forall i \notin S \right\}. \tag{EC.11}$$

Especially, for any given set S, the function $\Lambda(S)$ can be computed exactly since it is the trust region subproblem (TRS) that has been studied extensively in literature. Based on this observation, we propose a greedy algorithm for finding an approximate optimum solution in Algorithm 5. A key to the greedy algorithm is to compute $\Lambda(S)$. Here we implement the solver of this task outlined in Adachi et al. [1]. See Algorithm 7 for details. Also, we present a local search heuristic that can be more computationally expensive but has the potential to obtain better approximation solutions in Algorithm 6. We leverage the theoretical analysis and numerical study of this local search heuristic in future work.

Algorithm 5 Greedy algorithm for solving (Quad-MMD)

Require: Problem parameters (A, t), integer $d \in [D]$.

Let $\widehat{S}_{\text{greedy}} = \emptyset$ be the chosen set.
for $\ell = 1, 2, \ldots, d$ **do**
  Compute

$$j^* = \operatorname*{arg\,max}_{j\in[D]\setminus\widehat{S}_{\text{greedy}}} \Lambda(\widehat{S}_{\text{greedy}} \cup \{j\}).$$

  Add j^* to $\widehat{S}_{\text{greedy}}$.
end for
Return $\widehat{S}_{\text{greedy}}$.

Algorithm 6 Local search algorithm for solving (Quad-MMD)

Require: Problem parameters (A, t), integer $d \in [D]$, initial guess of a size-d subset $\widehat{S}$.

while There is still an improvement **do**
  for each pair $(i, j) \in \widehat{S} \times ([D] \setminus \widehat{S})$ **do**
    if $\Lambda(\widehat{S} \cup \{j\} \setminus \{i\}) > \Lambda(\widehat{S})$ **then**
      Update $\widehat{S} = \widehat{S} \cup \{j\} \setminus \{i\}$.
    end if
  end for
end while
Return $\widehat{S}$.

Algorithm 7 Eigenvalue decomposition algorithm for solving (EC.11)

Require: Problem parameters (A, t), set $S \subseteq [D]$ with $|S| = d$.

1: Formulate $A' = -A_{S,S}$ and $t' = -t_S$.

2: Formulate $2d \times 2d$ matrics

$$M_0 = \begin{pmatrix} -I_d & A' \\ A' & -t'(t')^{\mathrm{T}} \end{pmatrix}, \quad M_0 = \begin{pmatrix} 0 & I_d \\ I_d & 0 \end{pmatrix}.$$

3: Compute the rightmost eigenvalue λ^* of $M_0 + \lambda M_1$ and an eigenvalue $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ such that

$$M_0 \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = -\lambda^* M_1 \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

4: **if** $y_2^{\mathrm{T}} t' = 0$ **then**

5: Terminate and find solvers for the hard-case of TRS.

6: **else**

7: Obtain $p^* = -\mathrm{sign}(y_2^{\mathrm{T}} t') \frac{y_1}{\|y_1\|_2}$.

Return $\Lambda(S) = -(p^*)^{\mathrm{T}} A' p^* - (p^*)^{\mathrm{T}} t'$.

8: **end if**