Variable Selection for Kernel Two-Sample Tests

Jie Wang, Santanu S. Dey, Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

Funded by Office of Naval research (Grant number N000142212632), NSF CMMI-2112533, DMS-2134037

Question: How to Compare Two Samples

- **Given**: Samples from unknown distributions P and Q in $\mathbb{R}^D$.
- **Goal**:
  - Do P and Q differ?
  - Select d variables maximally distinguishing differences between P and Q!

$\sim P$

$\sim Q$

Maximum Mean Discrepancy (MMD)

- A kernel function $K(\cdot,\cdot)$ is called a positive semi-definite kernel if

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

- A positive semi-definite kernel K induces a unique RKHS $\mathcal{H}_K$.
- MMD statistic:

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

- Squared MMD statistic:

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

- Empirical MMD estimator:

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

MMD Variable Selection

- Pick the optimal variable selection z to maximize MMD:

$$\max_{z\in\mathcal{Z}} \quad S^2(\mathbf{x}^n,\mathbf{y}^m;K_z)$$

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

Statistical Performance Guarantees

Define the sample size $N = n \wedge m$ and

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

- Under null hypothesis $H_0: \mu=\nu$, with high probability,

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

- Under mild assumptions regarding μ and ν under H_1, it holds that

$$S(\mathbf{x}^n,\mathbf{y}^m;K_{\hat{z}}) \geq \Delta - O(1/\sqrt{N}),$$

where $\Delta > 0$ is a sufficiently large number.

- Example: Linear Kernel MMD. For $K_z(x,y) = \sum_{k\in[D]} z[k]x[k]y[k]$,

$$\max_{z\in\mathcal{Z}} \ a^{\mathrm{T}} z, \qquad a[k] = \left(\frac{1}{n}\sum_{i\in[n]} x_i[k] - \frac{1}{m}\sum_{j\in[m]} y_j[k]\right)^2.$$

Advantages: closed-form solution available!
Concerns: Only first-order moment condition is used!

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

Quadratic Kernel MMD

- For $K_z(x,y) = \left(\sum_{k\in[D]} z[k]x[k]y[k] + c\right)^2$, reduces to MIQP:

$$\max_{z\in\mathbb{R}^D} \left\{ S^2(\mathbf{x}^n,\mathbf{y}^m;K_z) = z^{\mathrm{T}}Az + z^{\mathrm{T}}t : \ \|z\|_2 = 1, \|z\|_0 = d \right\}.$$

When $t=0$, standard **sparse PCA** formulation (Li and Xie, 2020).

- Combinatorial formulation:

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

For fixed set S, it reduces to the **trust-region subproblem** (TRS) that is efficiently solvable.

Mixed-integer SDP reformulation

The Q-MMD optimization is equivalent to

$$\begin{aligned} \max_{Z\in\mathbb{S}^+_{D+1}, q\in\mathcal{Q}} \quad & \langle \tilde{A}, Z\rangle \\ \text{s.t.} \quad & Z_{i,i} \leq q[i], \quad i\in[D], \\ & Z_{0,0} = 1, \mathrm{Tr}(Z) = 2, \end{aligned}$$

where the set $\mathcal{Q} = \{q\in\{0,1\}^D : \ \sum_{k\in[D]} q_i = d\},$. It further admits two valid inequalities:

$$\begin{aligned} \sum_{j\in[D]} |Z_{i,j}| &\leq \sqrt{d}q[i], \quad \forall i\in[D] \\ |Z_{i,j}| &\leq M_{i,j}q[i], \quad \forall i,j\in[D] \end{aligned}$$

where $M_{i,j} = 1$ for $i=j$ and otherwise $M_{i,j} = 1/2$.

– Exact algorithm: cutting-plane algorithm;
– Approximation algorithm: convex relaxation of MISDP.

Performance Guarantees of Convex Relaxation

$$\mathrm{optval(MISDP)} \leq \mathrm{optval(SDP)} \leq \|t\|_2 + \min\left\{ D/d\cdot\mathrm{optval(MISDP)}, d\cdot\mathrm{optval(MISDP)} - \min_k |t[k]| \right\}.$$

- Population quadratic MMD statistic:

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

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

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

Only first- and second-order moment conditions are used for Q-MMD variable selection!

Gaussian Kernel MMD

- Define $Z := zz^{\mathrm{T}} \in \mathbb{S}^+_D$ and $M_{x,y} := \frac{1}{2\gamma}(x-y)(x-y)^{\mathrm{T}} \in \mathbb{S}^+_D$.
- Consider Gaussian kernel

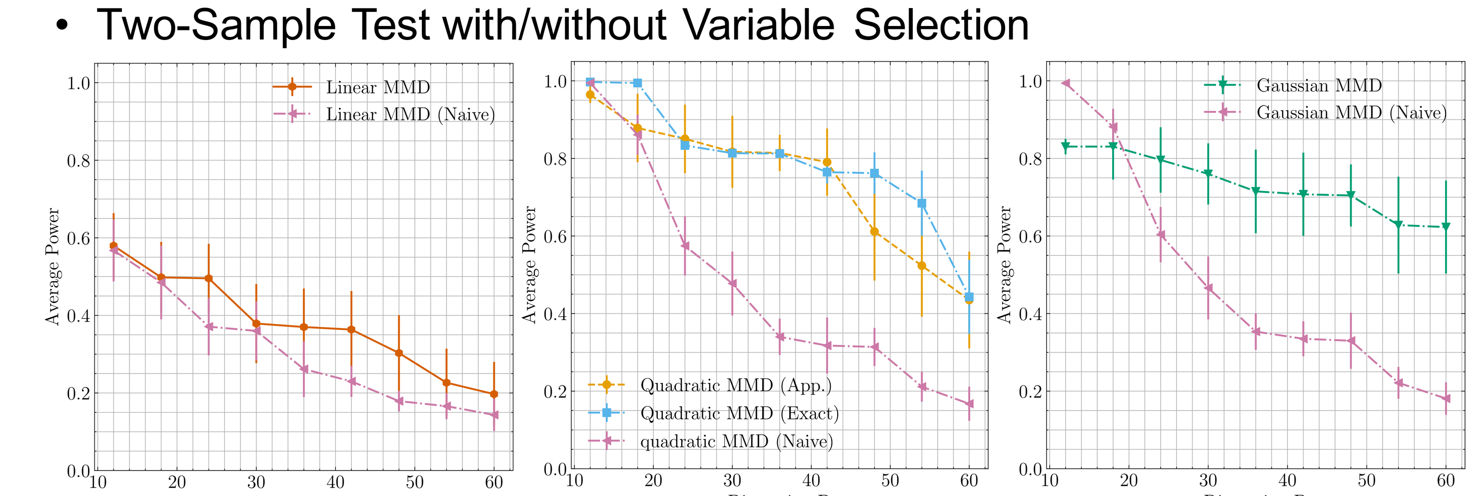
$$K_z(x,y) = \exp\left[-\frac{\left(\sum_{k\in[D]} z[k](x[k]-y[k])\right)^2}{2\gamma} \right] = \exp\left(-\frac{\langle Z, M_{x,y}\rangle}{2\gamma} \right)$$

- Rank and ℓ_0-norm constraint optimization:

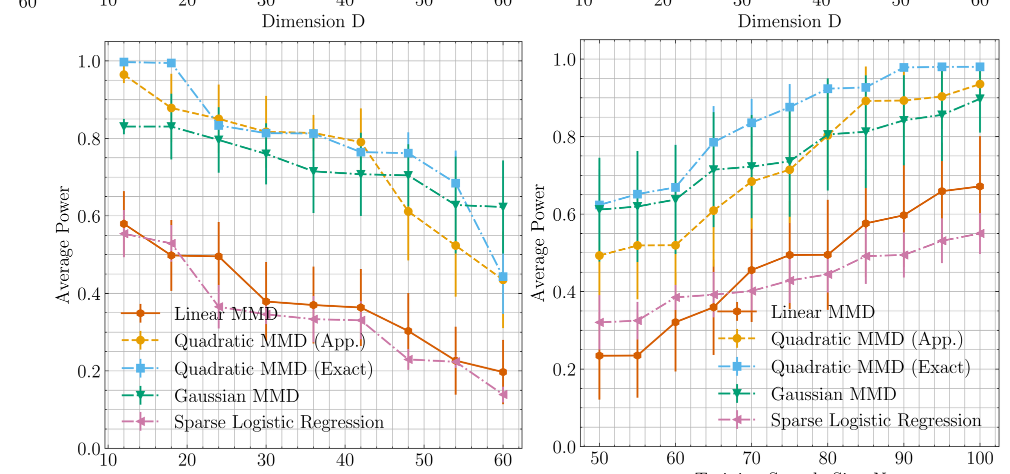
$$\begin{aligned} \min_{Z\in\mathbb{S}^+_D:\mathrm{Tr}(Z)=1} \quad & \frac{2\sum e^{-\langle Z, M_{x_i,y_j}\rangle}}{mn} - \frac{\sum e^{-\langle Z, M_{x_i,x_j}\rangle}}{n^2} - \frac{\sum e^{-\langle Z, M_{y_i,y_j}\rangle}}{m^2} \\ \text{s.t.} \quad & \|Z\|_0 \leq d^2, \mathrm{rank}(Z) = 1, \end{aligned}$$

Numerical Study

- Two-Sample Test with/without Variable Selection



- Two-Sample Test with Synthetic Dataset



- Two-Sample Test with Large-Scale Dataset
- $D = 500$, $d^* = 20$
- Non-discovery proportion : $\frac{|I^*\setminus I|}{|I^*|}$
- False-discovery proportion : $\frac{|I\setminus I^*|}{|I|}$