Variable Selection for Kernel Two-Sample Tests

Jie Wang, Santanu S. Dey, Yao Xie

Georgia Tech ISyE H. Milton Stewart School of Industrial and Systems Engineering

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 $\mathbb{R}^D$.

 $\sim P$
- Does P and Q differ?

 $\sim Q$
- Select d variables that maximally distinguish differences between P and Q

Maximum Mean Discrepancy (MMD)

- Kernel function $K(\cdot,\cdot)$ is positive semi-definite (PSD) if

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

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

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

- 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:

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

MMD Variable Selection

- Pick the optimal variable selection z to maximize MMD:

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

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.

- 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!
Only mean condition is used: $\overline{x} = \mathbb{E}[\mu], \overline{y} = \mathbb{E}[\nu]$,

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

Quadratic Kernel MMD

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

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

- 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 **trust-region subproblem**.

Mixed-integer SDP reformulation

The Q-MMD optimization is equivalent to

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

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

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

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

– **Exact algorithm:** cutting-plane algorithm;

– **Approximation algorithm:**

(I) Return the best over all d-sparse truncation of columns of A and all basis vectors:

$$V_{(\mathrm{I})} \geq \mathrm{optval}(\mathrm{MIQP})/\sqrt{d} - 2\|t\|_{(d+1)}.$$

(II) Solve the problem by dropping ℓ_0-norm constraint and return its d-sparse truncation:

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

- 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

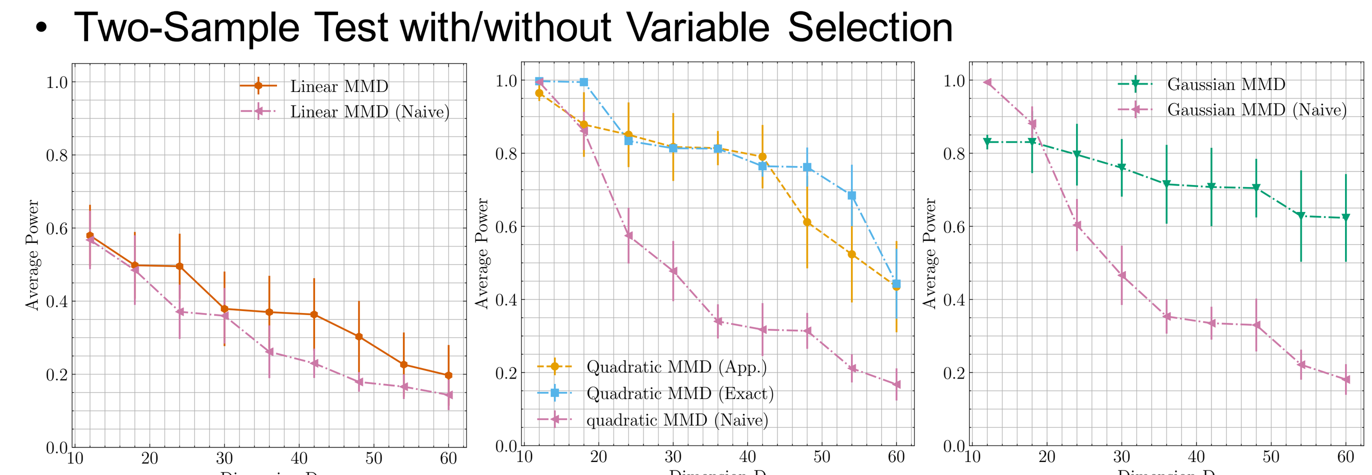
$$(\mathcal{A}(\mu,\nu))_{k_1,k_2} = (\mathbb{E}_{x\sim\mu}[x[k_1]x[k_2]] - \mathbb{E}_{y\sim\nu}[y[k_1]y[k_2]])^2$$

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

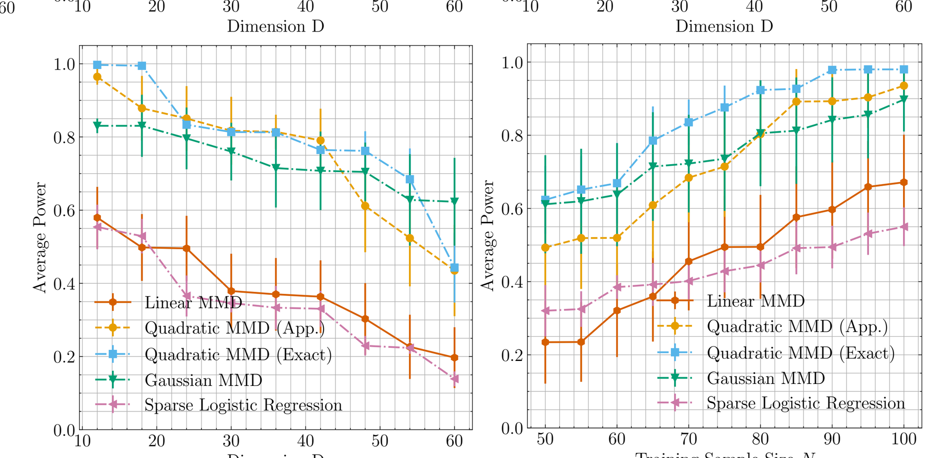
$$\mathcal{T}(\mu,\nu)[k] = 2c\left(\mathbb{E}_{x\sim\mu}[x[k]] - \mathbb{E}_{y\sim\nu}[y[k]]\right)^2.$$

Only 1st and 2nd-order moment conditions are used.

- 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 : $\frac{|I\setminus I^*|}{|I|}$