Statistical and Computational Guarantees of Kernel Max-Sliced Wasserstein Distances

Jie Wang
Georgia Institute of Technology

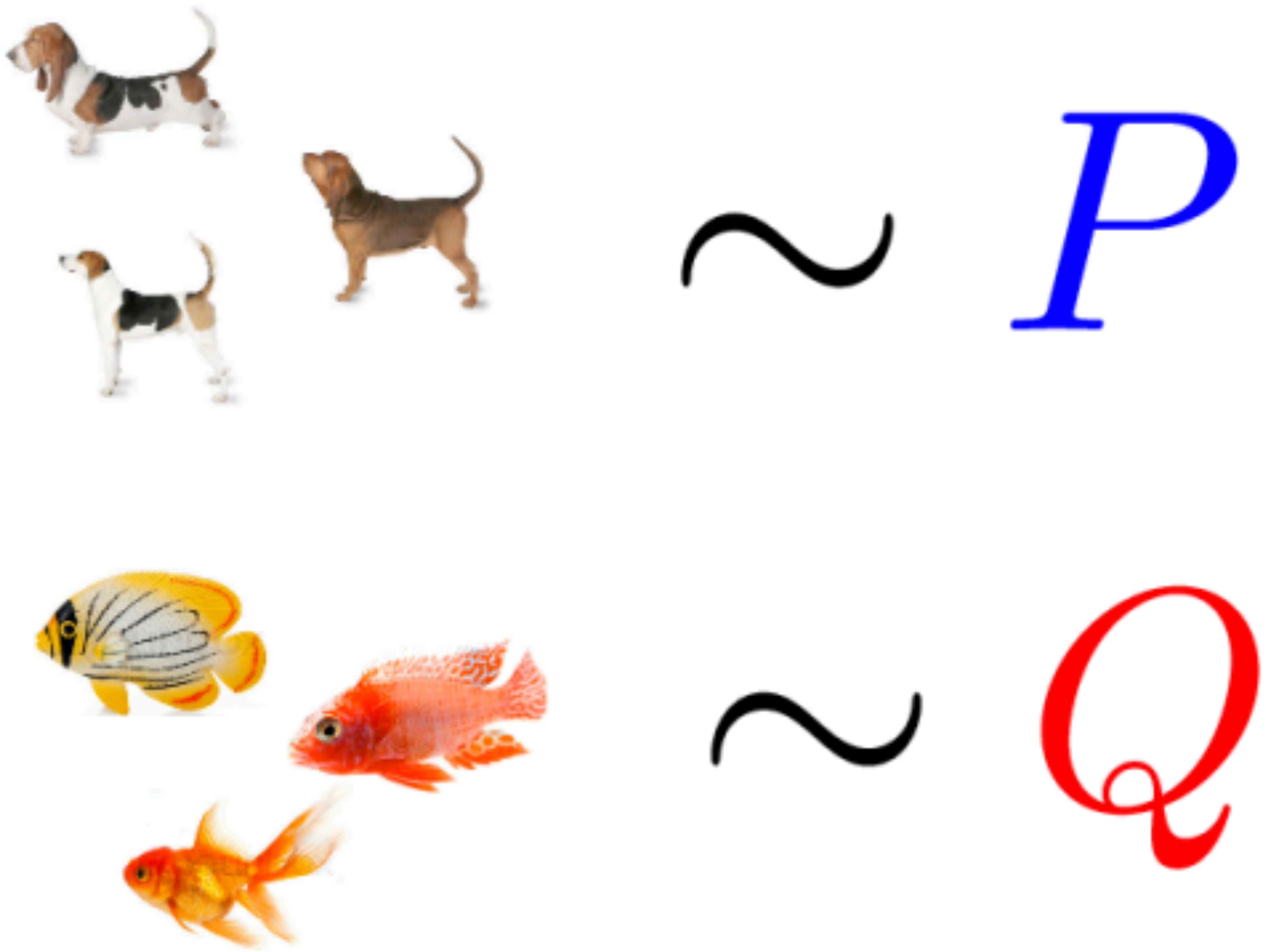
Data Mining Best Student Paper Competition

Joint work with March Boedihardjo (Michigan State) and Yao Xie (Georgia Tech)

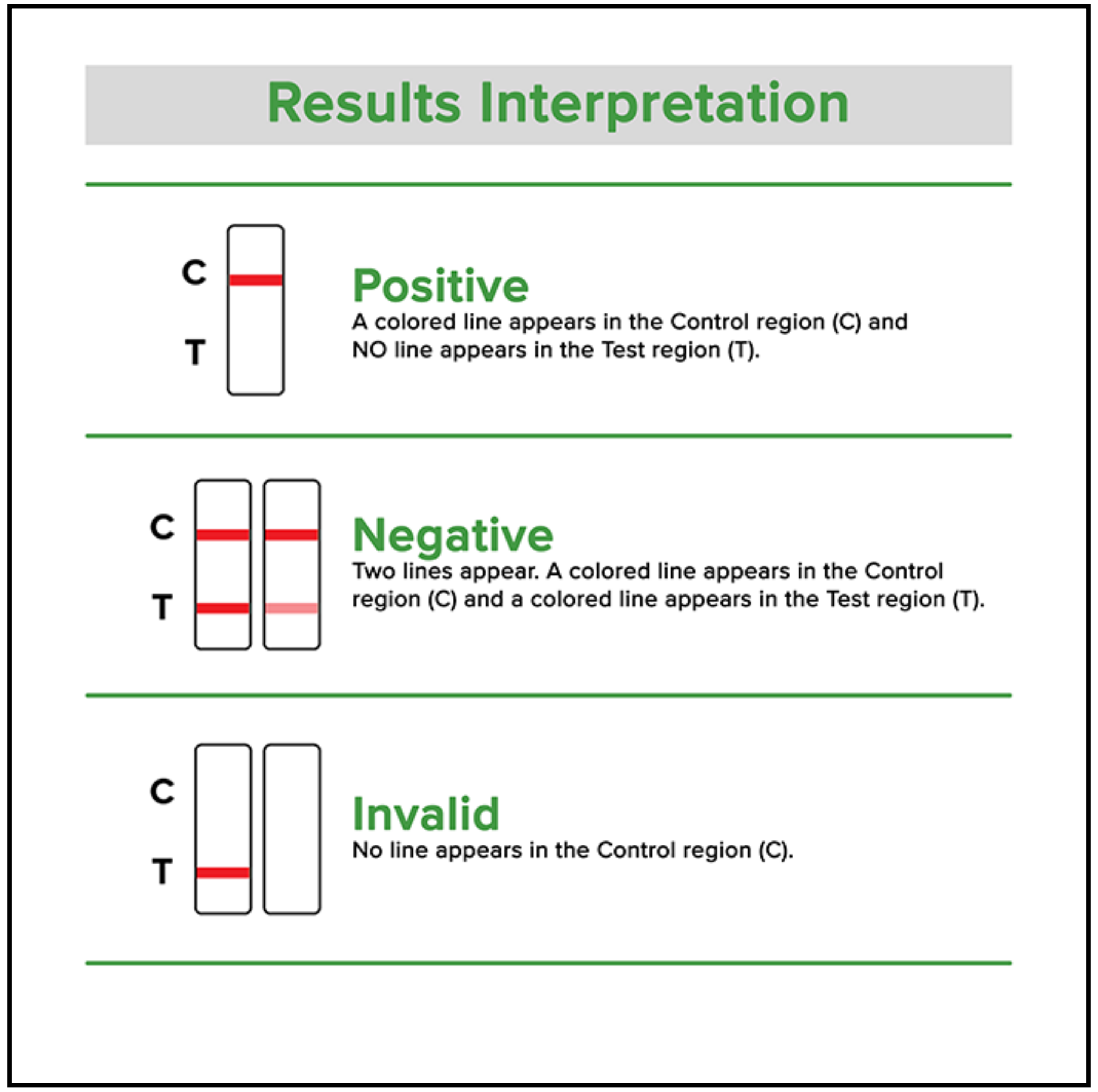
0. Introduction

Question: How to Compare Two Samples

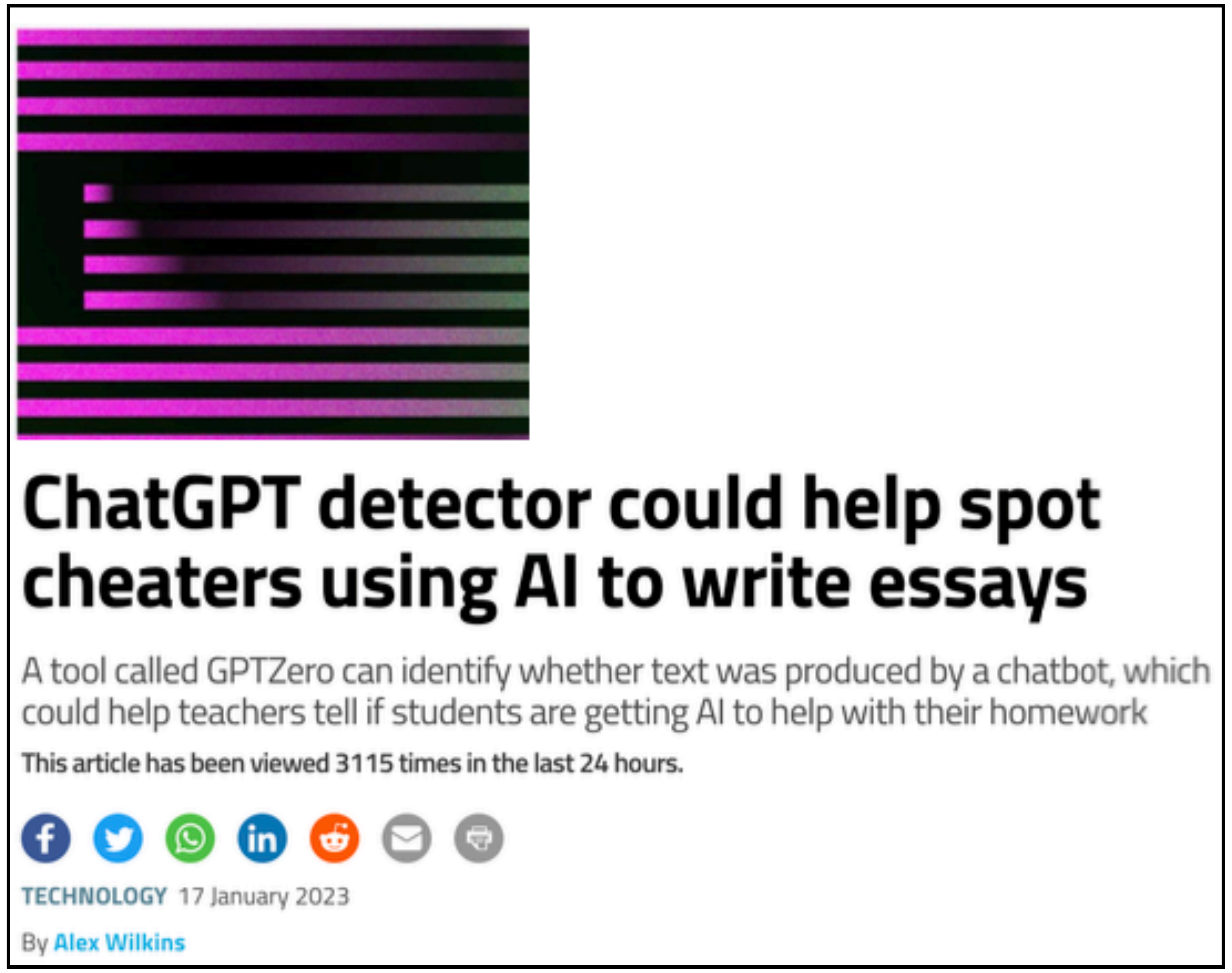
- **Given**: Two high-dimensional data samples from unknown distributions P and Q
- **Goal**: Does P and Q differ?



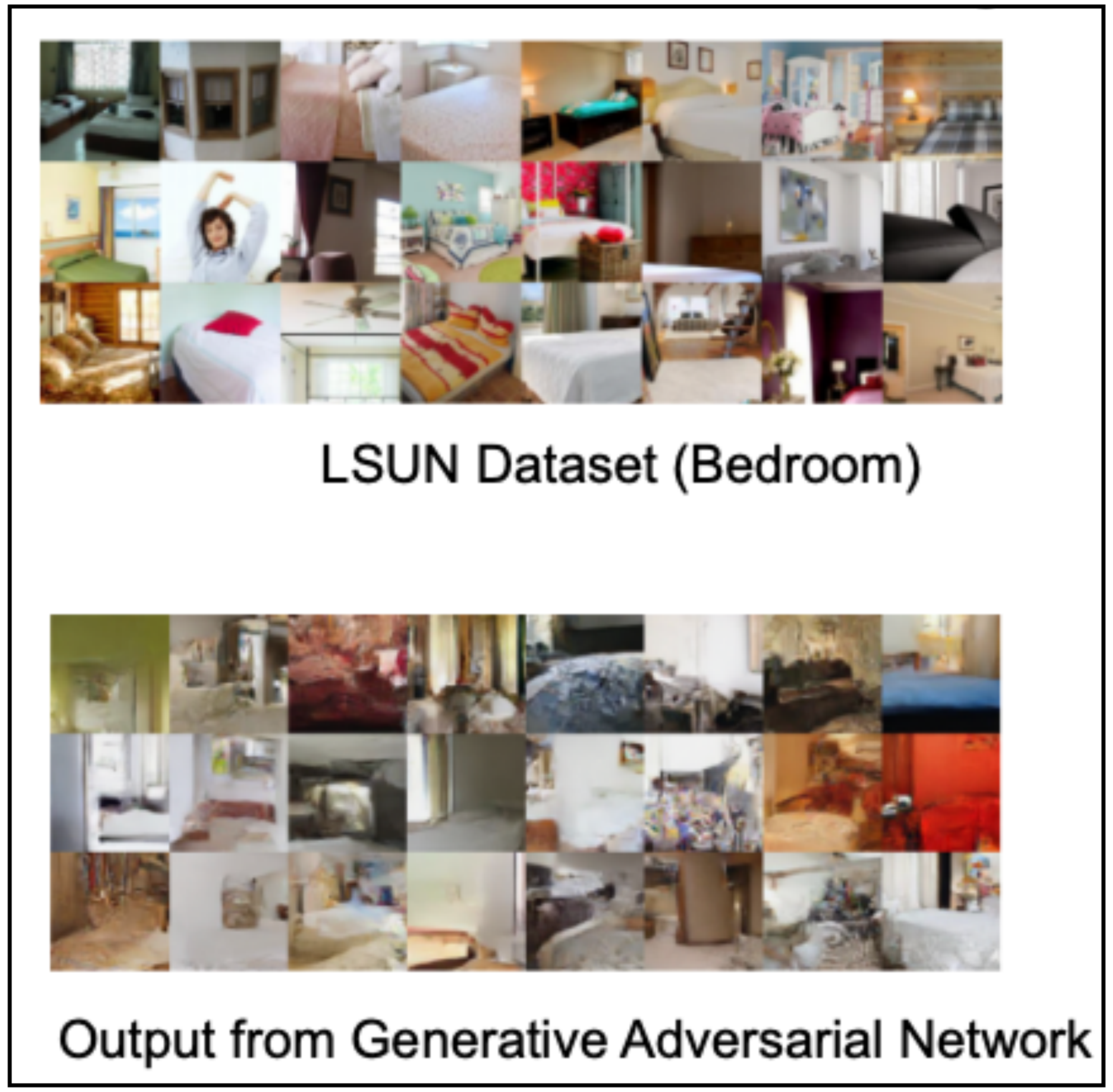
Two-Sample Test is Fundamental in Practice



Covid-19 Test

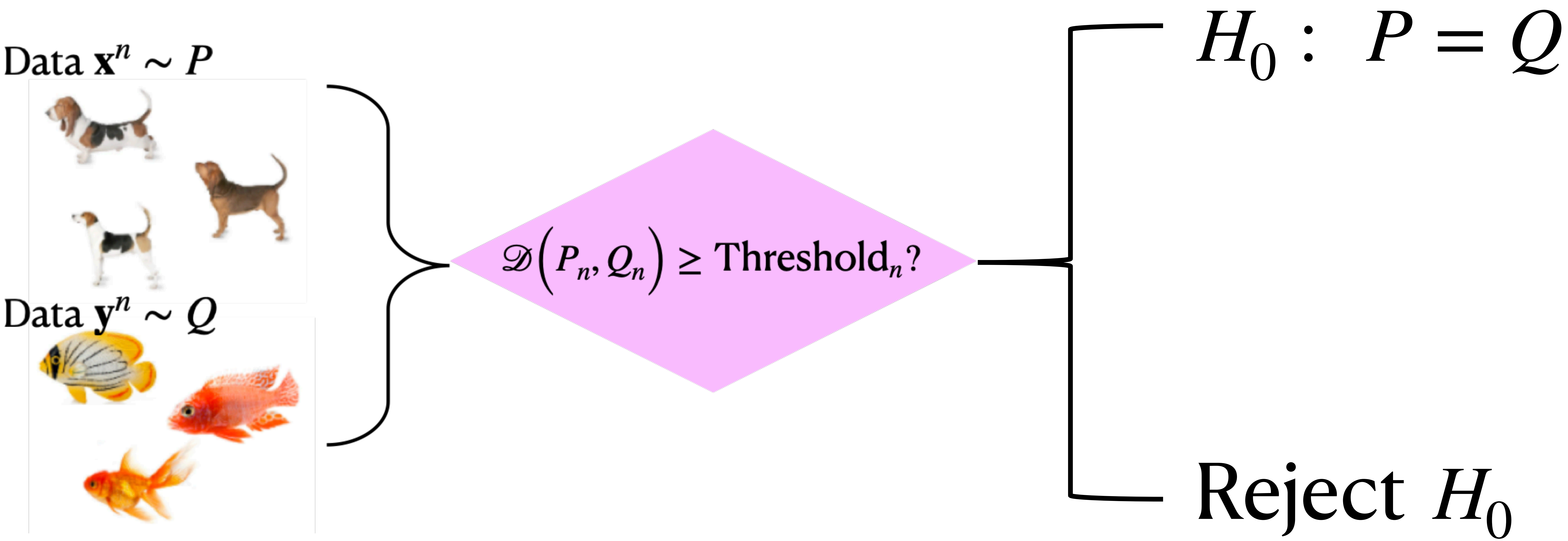


Goodness of Fit Test



Model Criticism

Non-parametric Two-Sample Test



Non-parametric Two-Sample Test

Data $\mathbf{x}^n \sim P$

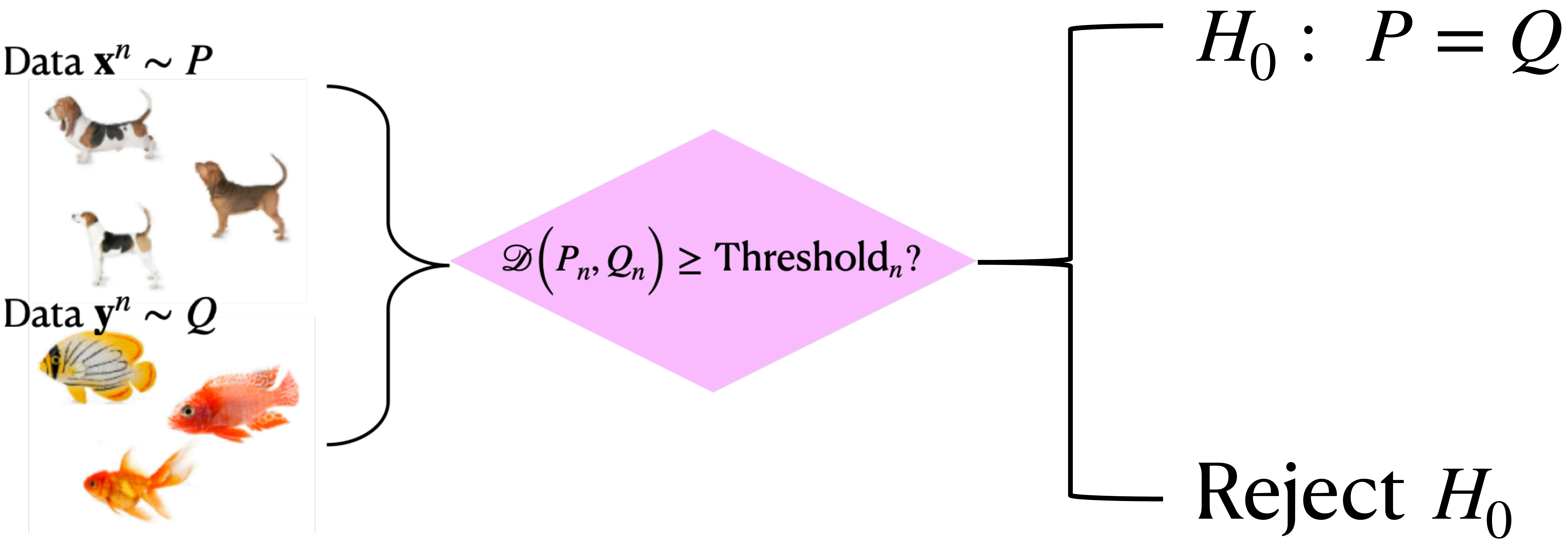
Data $\mathbf{y}^n \sim Q$

$\mathscr{D}\left(P_n, Q_n\right) \geq \text{Threshold}_n?$

$H_0: \; P = Q$

Reject H_0

Non-parametric Two-Sample Test



Goal: Develop an ***effective***, ***non-parametric*** metric $\mathscr{D}(\cdot\,,\cdot)$ to ***interpretably*** characterize differences between **high-dimensional** distributions

Wasserstein Distance

$$\mathbf{W}(P, Q) = \min_{\gamma} \left\{ \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)] : \quad \gamma \text{ has marginal distributions } P \text{ and } Q \right\}$$

- **Pros**: Flexible, non-parametric, incorporate geometric properties
- **Cons**: Testing power degrades in the rate of $O(n^{-1/d})$, curse of dimension!

[Ramdas A, 2017], [Fournier N, 2015]

Wasserstein Distance

$$\mathbf{W}(P, Q) = \min_{\gamma} \left\{ \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)] : \quad \gamma \text{ has marginal distributions } P \text{ and } Q \right\}$$

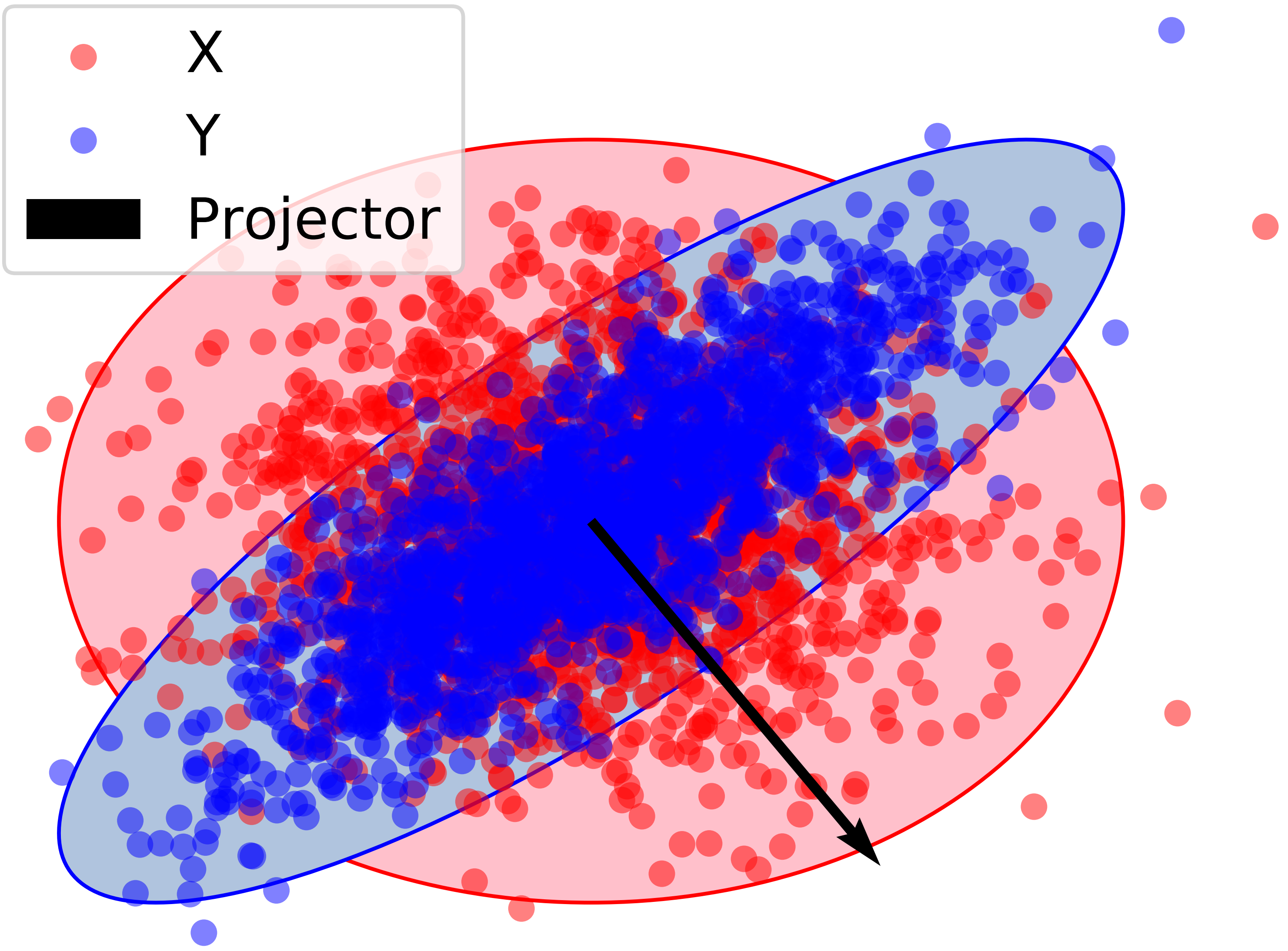
- **Pros**: Flexible, non-parametric, incorporate geometric properties
- **Cons**: Testing power degrades in the rate of $O(n^{-1/d})$, curse of dimension!

[Ramdas A, 2017], [Fournier N, 2015]

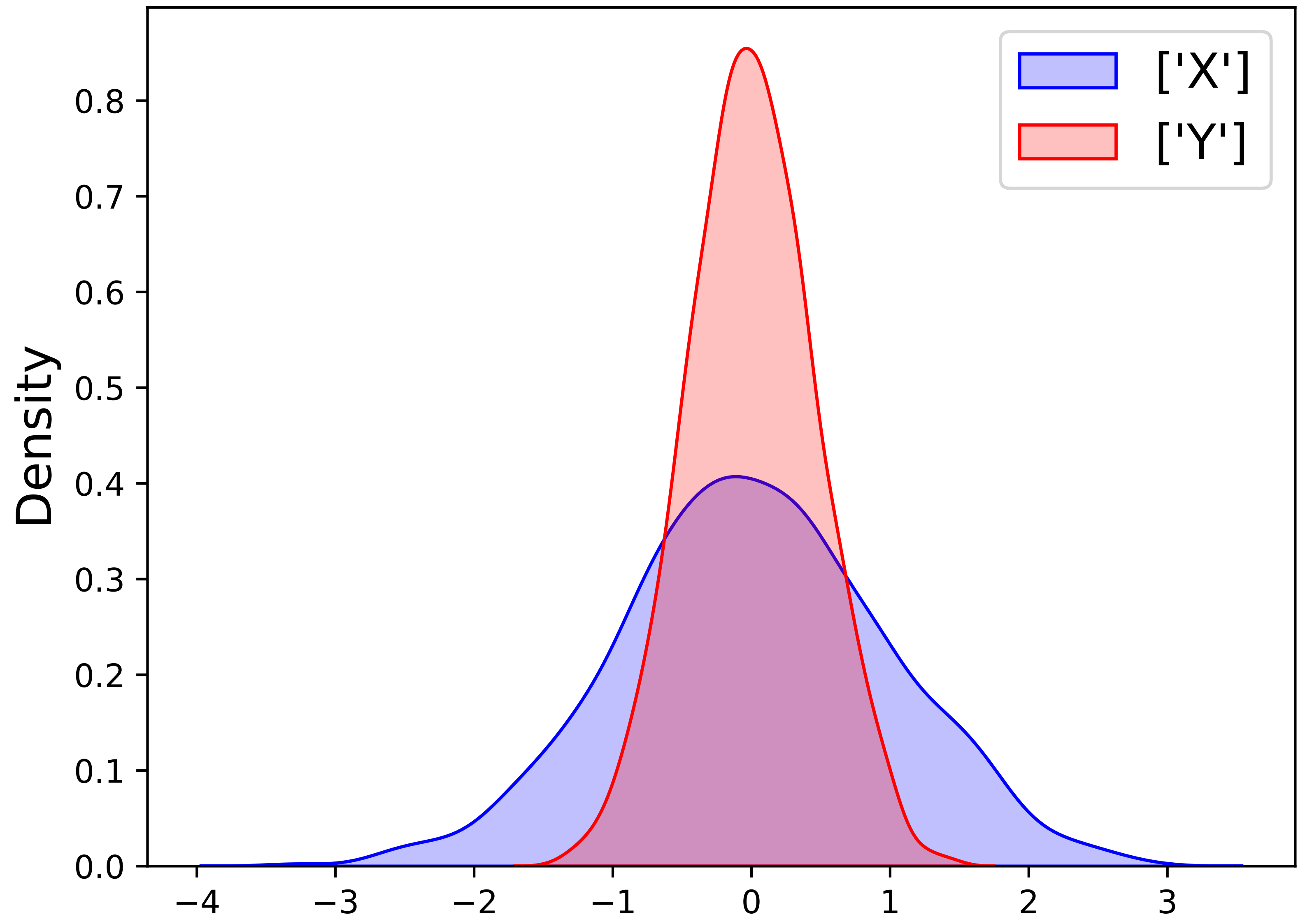
Max-Sliced Wasserstein Distance

$$\mathbf{MS}(P, Q) = \max_{v \in \mathbb{R}^d, \|v\|_2 = 1} \mathbf{W}(v_\# P, v_\# Q)$$

$v_\# P$ represents the distribution of P by the **linear projection** along the direction v



Scatter plot for 2-dimensional Gaussian



Density plot for linearly projected samples

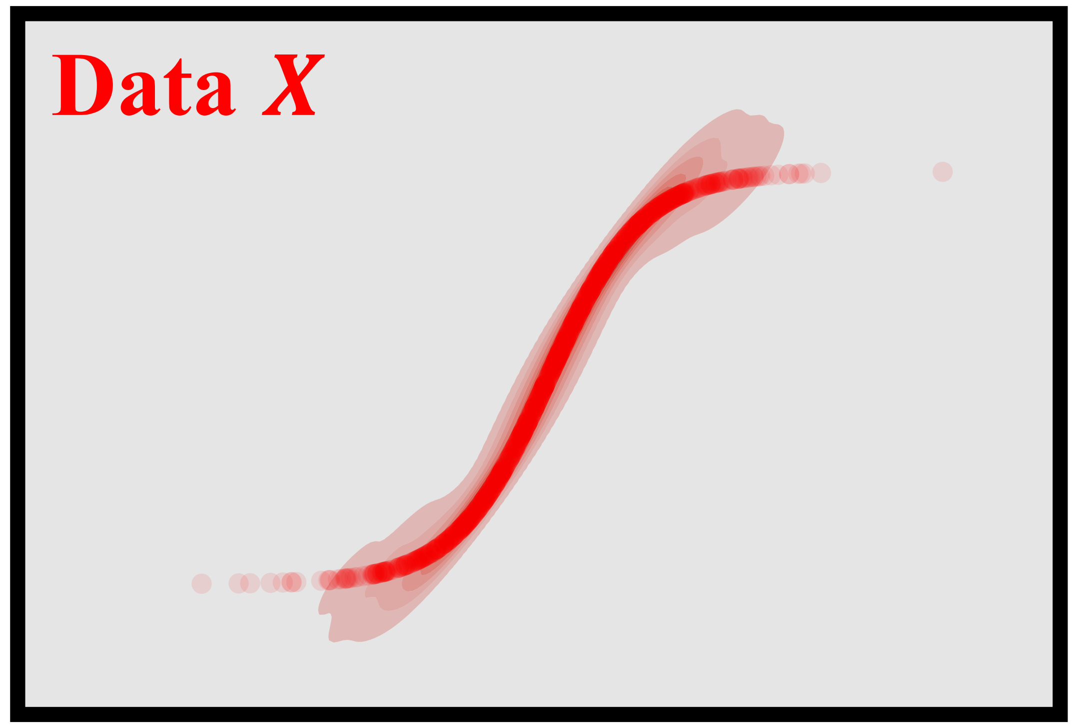
- **Pros**: Powerful for linearly separable data
- **Cons**: Performance degrades if data are nonlinearly separable

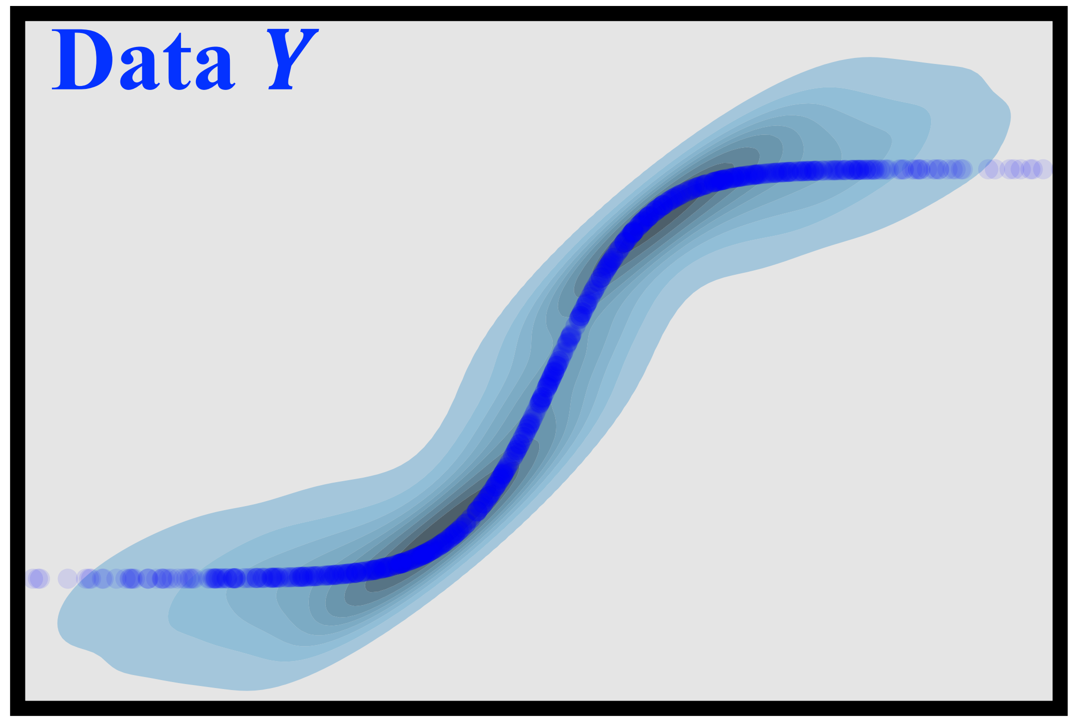
[Deshpande I et al, 2019], [Wang J, 2021]

Kernel Max-Sliced Wasserstein Distance

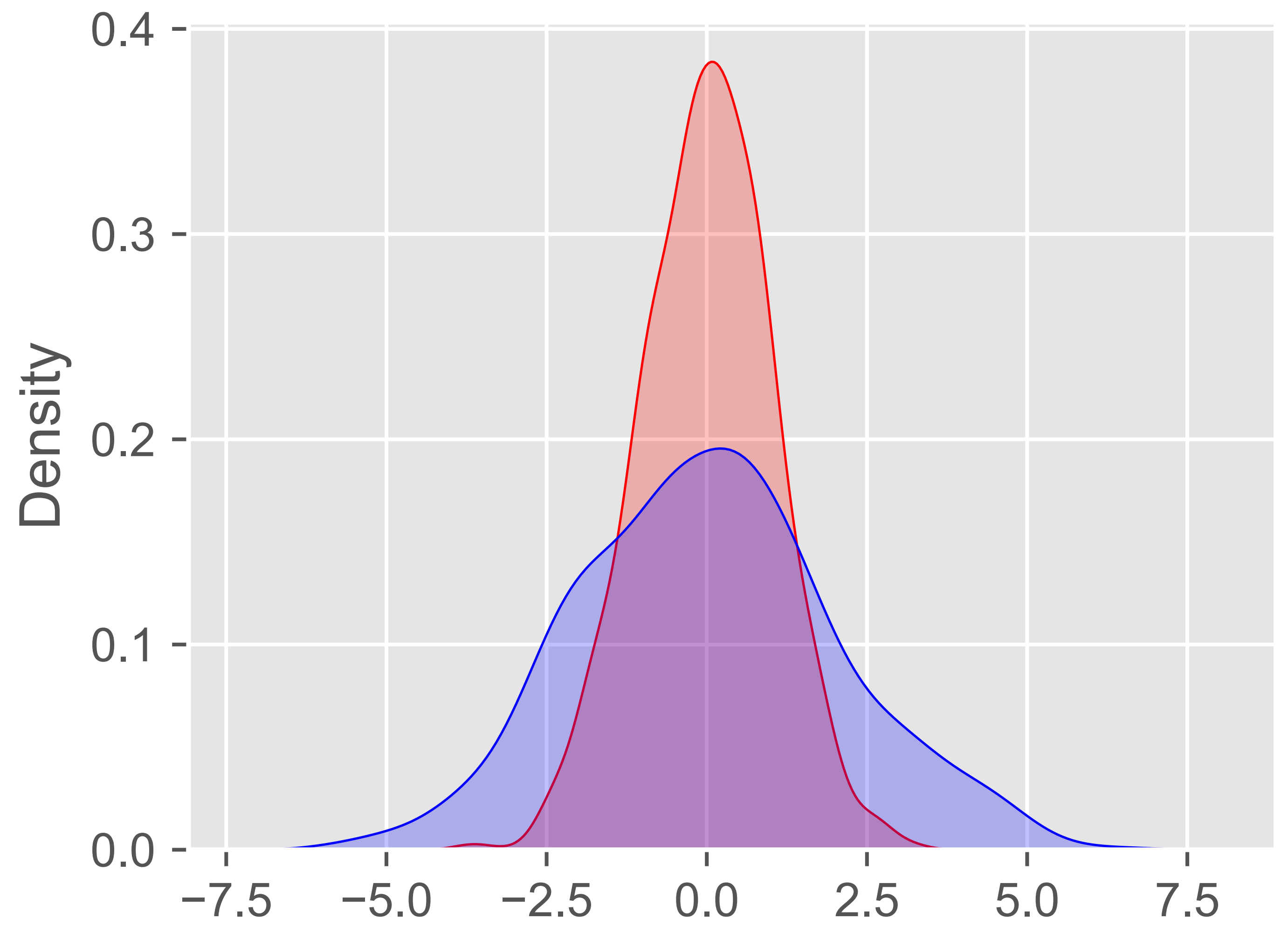
$$\mathbf{KMS}(P, Q) = \max_{f \in \mathscr{F}} \ \mathbf{W}(f_\# P, f_\# Q)$$

- $f_\# P$: push forward distribution of P under **nonlinear projector** f
- $\mathscr{F} = \left\{ f \in \mathscr{H} : \ \|f\|_{\mathscr{H}} \leq 1 \right\}$: A unit ball of reproducing kernel Hilbert space (RKHS)





Nonlinear projector computed from **KMS**



Roadmap

$$\mathbf{KMS}(P, Q) = \max_{f \in \mathscr{H} :\ \|f\|_{\mathscr{H}} \leq 1} \mathbf{W}(f_\# P, f_\# Q)$$

1. Why does KMS excel in characterizing differences between high-dimensional distributions?

2. How can KMS be computed efficiently while ensuring performance guarantees?

3. Practical Applications of KMS

1. Statistical Guarantees of KMS

Nonlinear Projection Function in RKHS

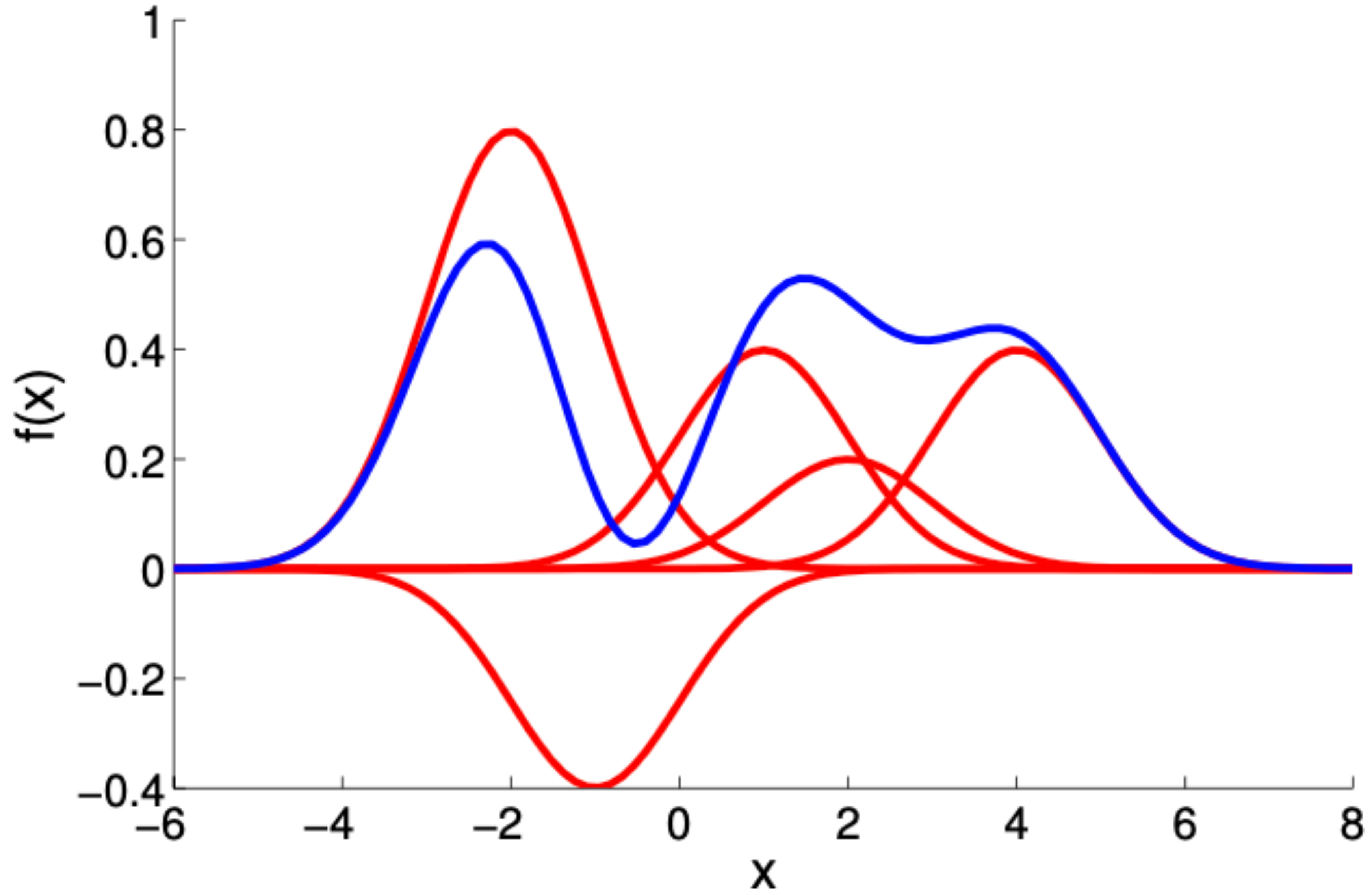
$$\mathbf{KMS}(P,Q) = \max_{f \in \mathcal{H}:\ \|f\|_{\mathcal{H}} \leq 1} \mathbf{W}(f_\# P, f_\# Q)$$

- $\mathcal{H}$: RKHS from $\mathbb{R}^d$ to $\mathbb{R}$ induced by **positive definite** kernel $k(\cdot,\cdot)$
- **Example:** $f \in \mathcal{H}$ if $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$ for arbitrary $m \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathbb{R}^d$
- $\mathbf{KMS}(P,Q) = 0$ iff $P = Q$ when $k(\cdot,\cdot)$ is **universal**

Example:

$k(x,y) = e^{-\|x-y\|_2^2/\sigma^2}$

$k(x,y) = e^{-\|x-y\|/\sigma}$



Finite-Sample Guarantees

$$\mathbf{KMS}(P, Q) = \max_{f \in \mathscr{H}:\ \|f\|_{\mathscr{H}} \le 1} \mathbf{W}(f_\# P, f_\# Q)$$

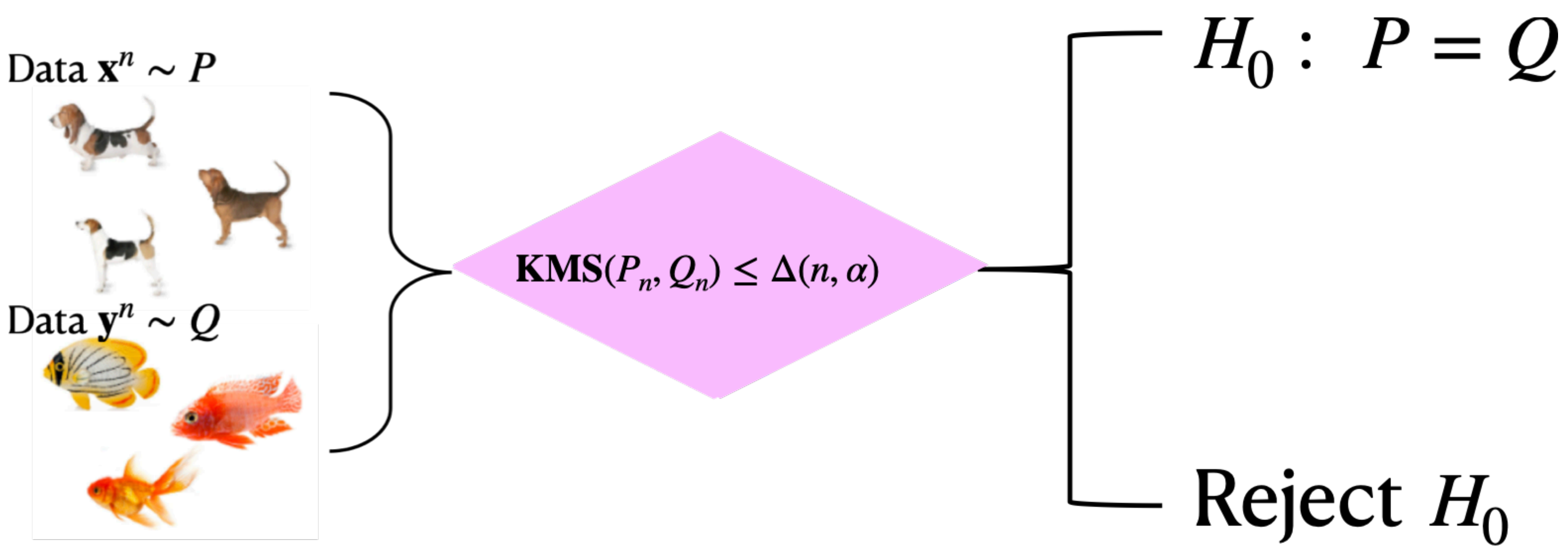
Theorem (Informal). **Assume** $k(x, x) \le A, \forall x$. With high probability,

$$\mathbf{KMS}(P, P_n) = O(n^{-1/2})\,.$$

- $O(\cdot)$ hides constant depending on A
- P_n denotes the empirical distribution of n i.i.d. samples from P

- **KMS** breaks the curse of dimensionality of Wasserstein distance
- Free of distribution assumptions
    - (typically studied in MS distance [Sloan N et. al, 2022, Tianyi L et. al, 2021])

Applications to Two-Sample Testing



Theorem (Informal). Fix level $\alpha \in (0,1/2)$ and Specify threshold $\Delta(n, \alpha) = O(\sqrt{\log(1/\alpha)} \cdot n^{-1/2})$. Then:

- **Type-I Error** of KMS test is at most α;
- Under $H_1 : \ P \neq Q$, the **power** of KMS test is at least $1 - O(n^{-1/2})$.

2. Computational Guarantees of KMS

Computing KMS Between Datasets

$$\mathbf{KMS}(P_n, Q_n) = \max_{f\in\mathcal{H}:\ \|f\|_{\mathcal{H}}\leq 1} \left\{ \min_{\pi\in\Gamma} \sum_{i,j=1}^{n} \pi_{i,j}\|x_i - y_j\|_2^2 \right\}$$

- $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \qquad Q_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$
- $\Gamma_n = \left\{\pi \in \mathbb{R}_+^{n\times n} : \ \sum_{i=1}^{n}\pi_{i,j} = \frac{1}{n}, \sum_{j=1}^{n}\pi_{i,j} = \frac{1}{n}\right\}$

Computing KMS Between Datasets

$$\mathbf{KMS}(P_n, Q_n) = \max_{f \in \mathscr{H}:\ \|f\|_{\mathscr{H}} \leq 1} \left\{ \min_{\pi \in \Gamma} \sum_{i,j=1}^{n} \pi_{i,j} \|x_i - y_j\|_2^2 \right\}$$

- $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}, \qquad Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- $\Gamma_n = \left\{ \pi \in \mathbb{R}_+^{n \times n} : \ \sum_{i=1}^{n} \pi_{i,j} = \frac{1}{n}, \sum_{j=1}^{n} \pi_{i,j} = \frac{1}{n} \right\}$

Theorem. There exists an optimal solution $\hat{f}$ such that

$$\hat{f}(z) = \sum_{i=1}^{n} k(z, x_i) a_{x,i} - \sum_{j=1}^{n} k(z, y_j) a_{y,j}$$

where $a_{x,i}, a_{y,j}$ are coefficients to be determined.

Enables finite-dimensional reformulation

Finite-Dimensional Reformulation

$$\mathbf{KMS}(P_n, Q_n) = \max_{\omega \in \mathbb{R}^{2n}:\ \|\omega\|_2 = 1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} (M_{i,j}^\top \omega)^2 \right\}$$

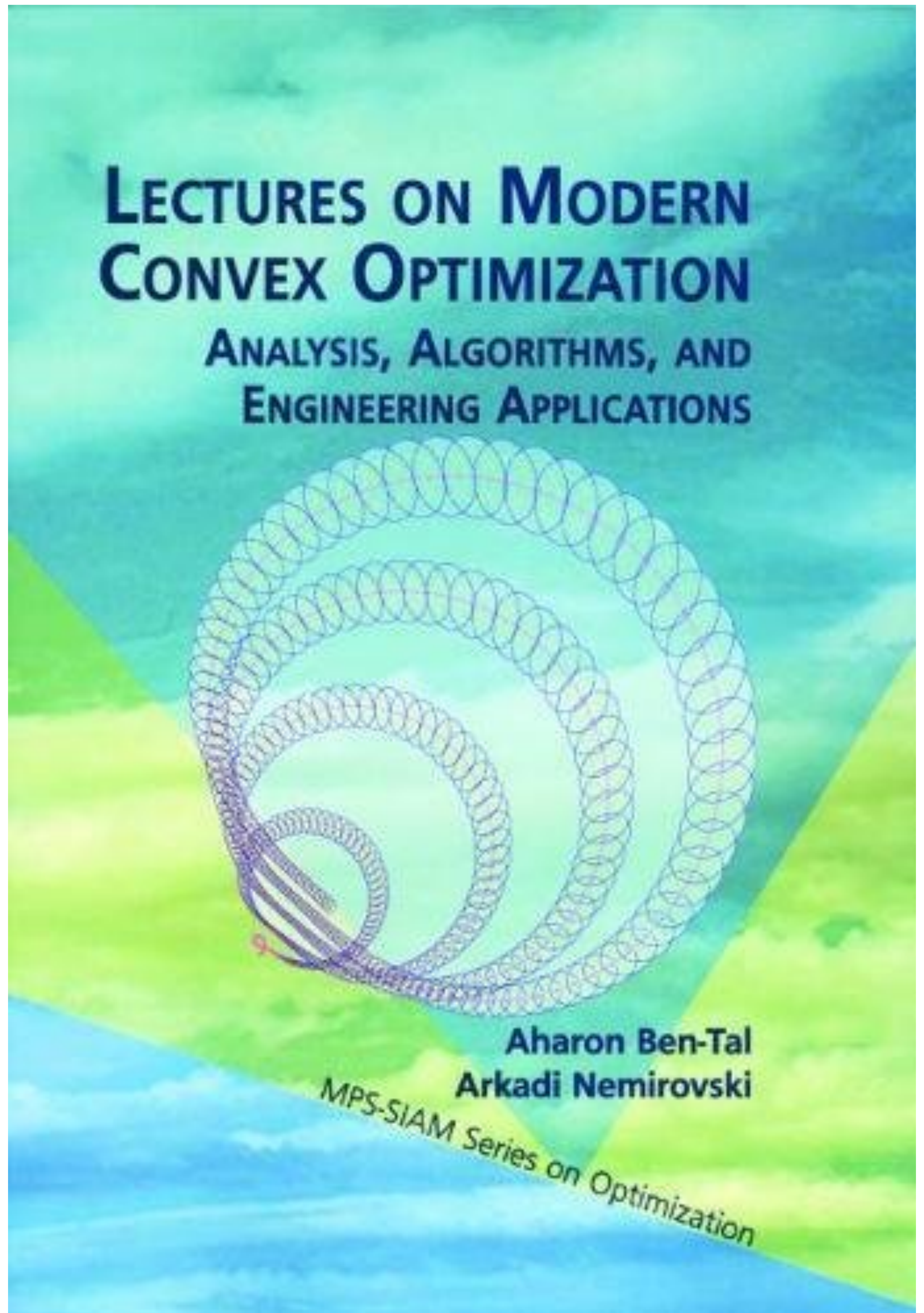
- $M_{i,j} \in \mathbb{R}^{2n}$: concatenation of kernel valued on data points
- Non-concave quadratic optimization problem

Theorem. $\mathrm{KMS}(P_n, Q_n)$ **is $\mathcal{NP}$-hard to compute**

Finite-Dimensional Reformulation

$$\mathbf{KMS}(P_n, Q_n) = \max_{\omega \in \mathbb{R}^{2n}:\ \|\omega\|_2 = 1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^\top, \omega\omega^\top \rangle \right\}$$

- **Approximation algorithm** using semidefinite relaxation (SDR):



$$S = \omega\omega^\top, \quad \omega \in \mathbb{R}^{2n}, \quad \|\omega\|_2 = 1$$

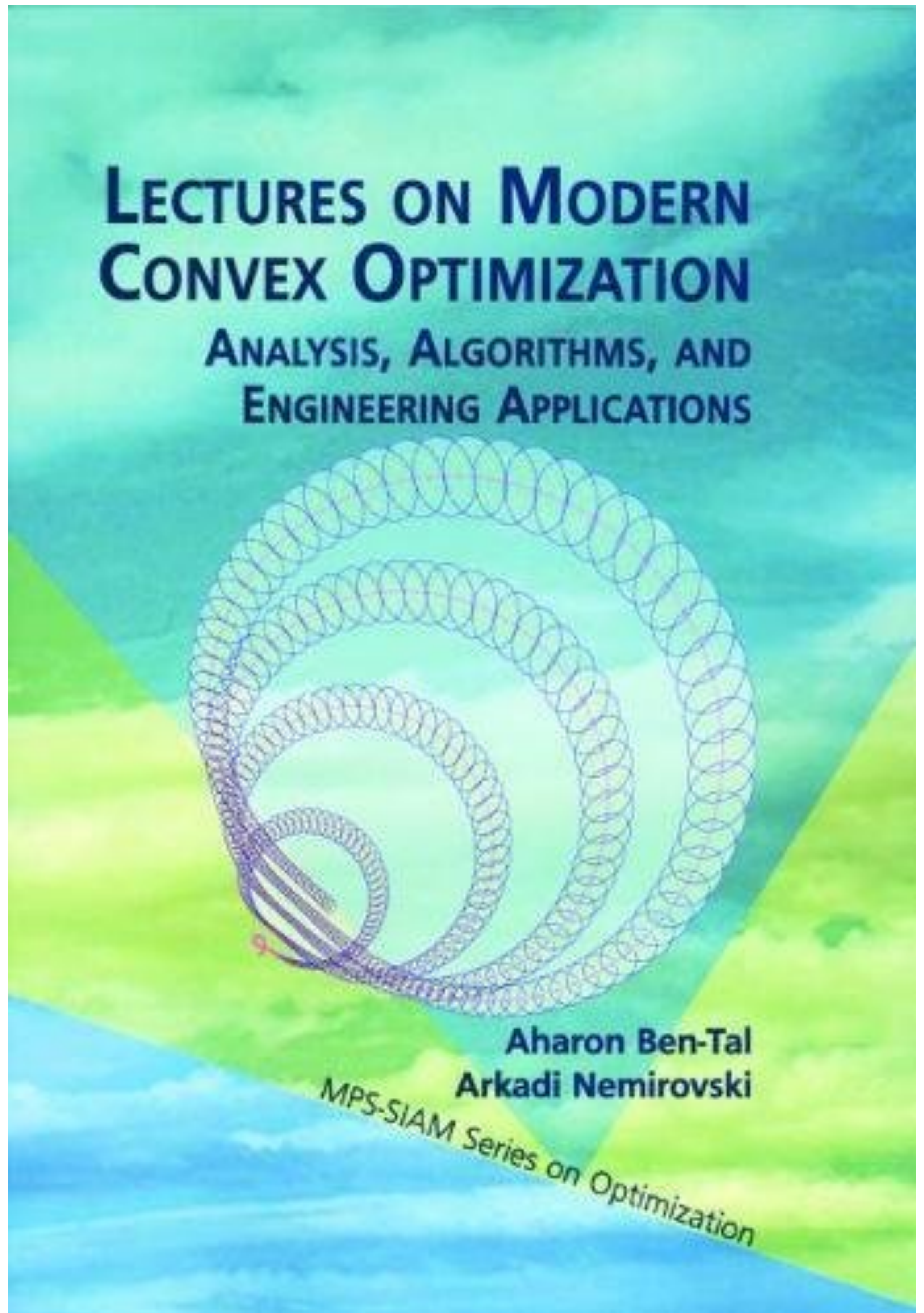
$$\Updownarrow$$

$$S \succeq 0, \quad \mathrm{Trace}(S) = 1, \quad \mathrm{rank}(S) = 1$$

Finite-Dimensional Reformulation

$$\mathbf{KMS}(P_n, Q_n) = \max_{\omega \in \mathbb{R}^{2n}:\ \|\omega\|_2 = 1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^\top, \omega\omega^\top \rangle \right\}$$

- **Approximation algorithm** using semidefinite relaxation (SDR):



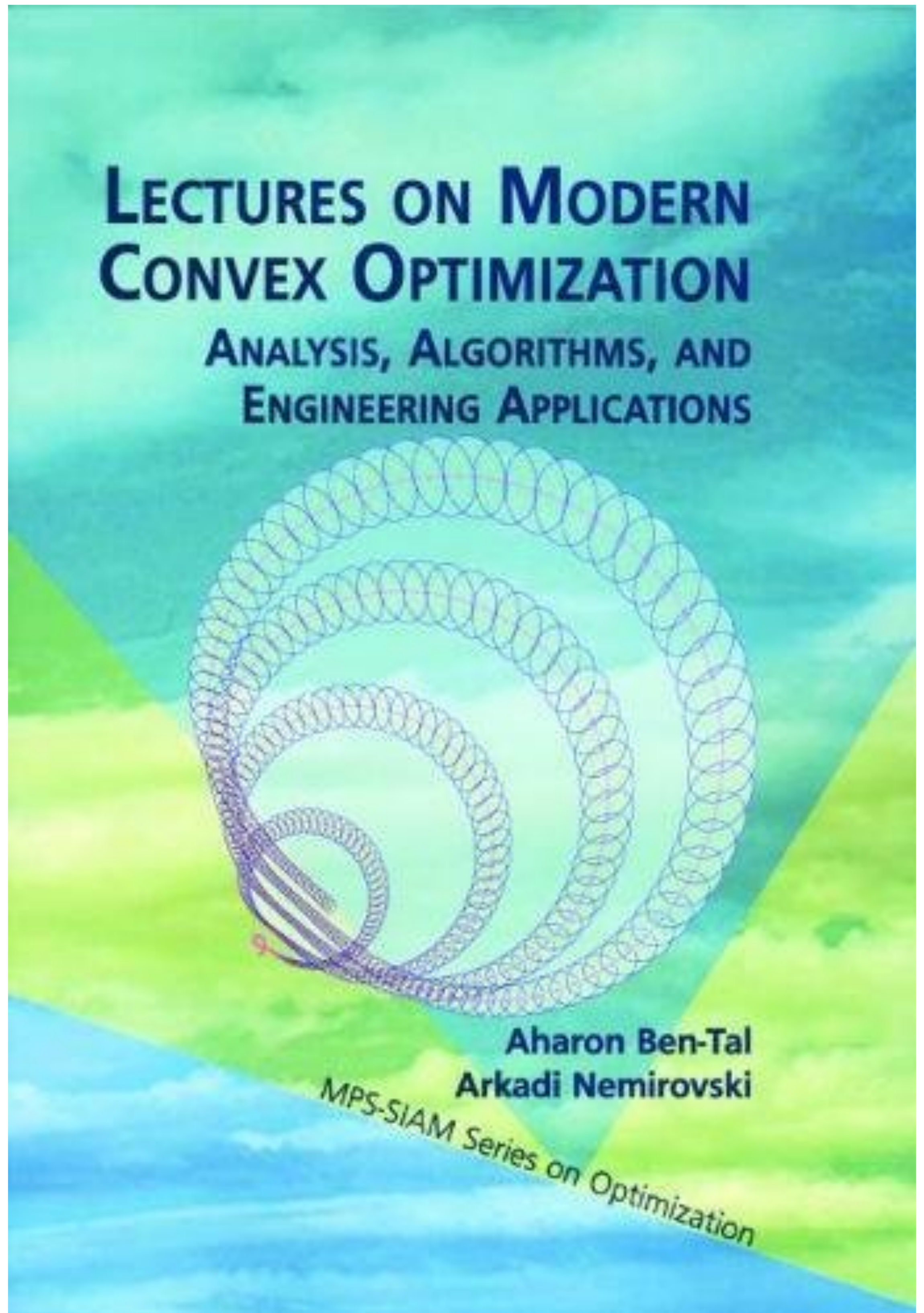
$$S = \omega\omega^\top, \quad \omega \in \mathbb{R}^{2n}, \quad \|\omega\|_2 = 1$$

$$S \succeq 0, \quad \mathrm{Trace}(S) = 1, \quad \mathrm{rank}(S) = 1$$

Finite-Dimensional Reformulation

$$\mathbf{KMS}(P_n, Q_n) = \max_{\omega \in \mathbb{R}^{2n}:\ \|\omega\|_2 = 1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\top}, \omega\omega^{\top} \rangle \right\}$$

- **Approximation algorithm** using semidefinite relaxation (SDR):



$$S = \omega\omega^{\top}, \quad \omega \in \mathbb{R}^{2n}, \quad \|\omega\|_2 = 1$$

$$S \succeq 0, \quad \mathrm{Trace}(S) = 1, \quad \cancel{\mathrm{rank}(S) = 1}$$

Semidefinite Relaxation (SDR)

$$\mathbf{KMS}(P_n, Q_n) = \max_{S \succeq 0, \mathrm{Trace}(S)=1, \mathrm{rank}(S)=1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\top}, S \rangle \right\}$$

$$\leq \max_{S \succeq 0, \mathrm{Trace}(S)=1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^{\top}, S \rangle \right\}$$

Theorem (Informal). Stochastic gradient method with biased oracles solves SDR up to δ optimality gap with operational complexity

$$O(n^2 (\log n)^{3/2} \delta^{-3})$$

Quality of Semidefinite Relaxation

- $$(\text{KMS}) = \max_{S \succeq 0, \text{Trace}(S)=1, \text{rank}(S)=1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^\top, S \rangle \right\}$$
- $$(\text{SDR}) = \max_{S \succeq 0, \text{Trace}(S)=1} \left\{ \min_{\pi \in \Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j} \langle M_{i,j} M_{i,j}^\top, S \rangle \right\}$$

Smaller rank of optimal solution yields better performance

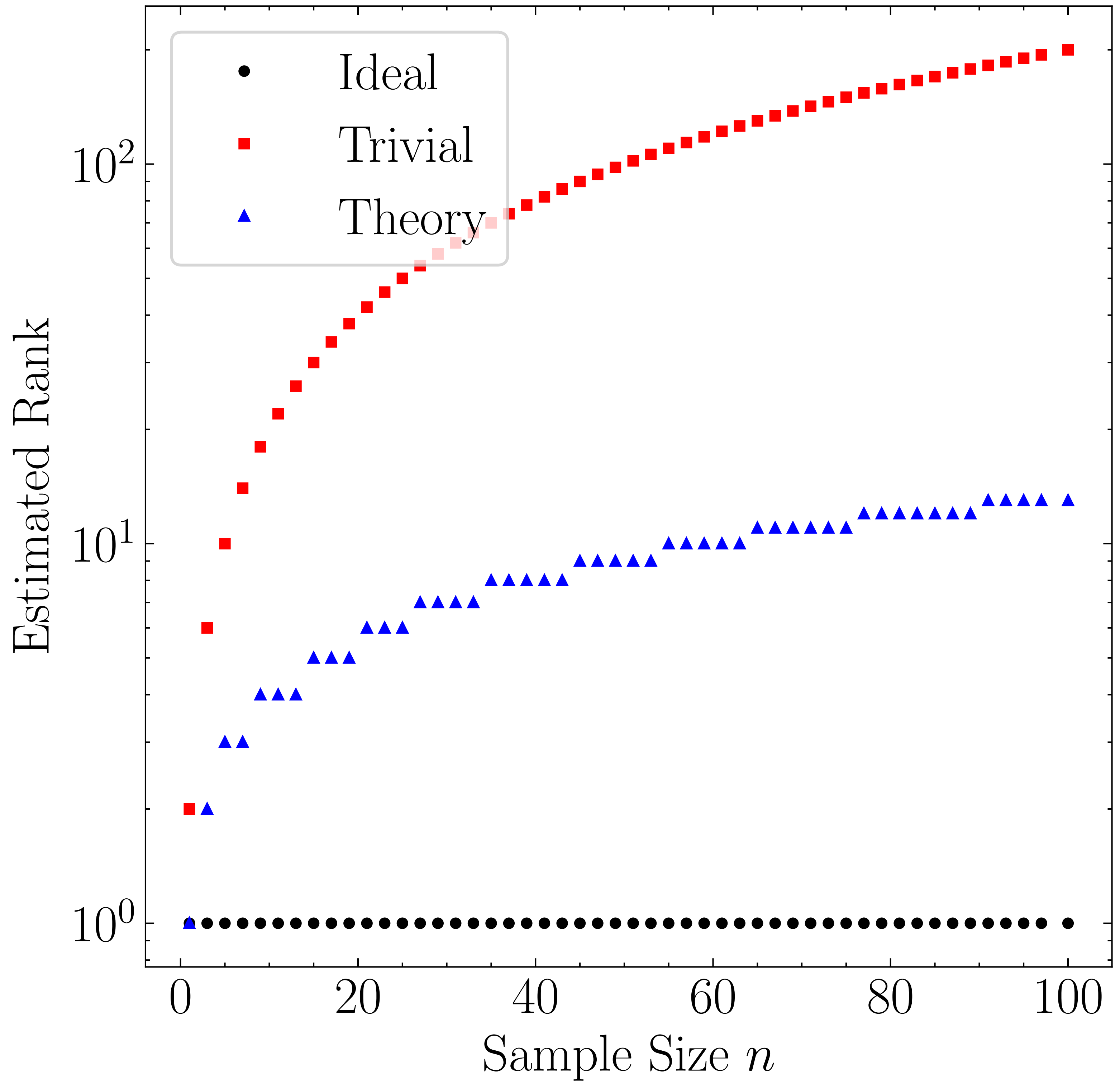
Quality of Semidefinite Relaxation

- $$(\text{KMS}) = \max_{S\succeq 0,\text{Trace}(S)=1,\text{rank}(S)=1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j}\langle M_{i,j}M_{i,j}^\top, S\rangle \right\}$$
- $$(\text{SDR}) = \max_{S\succeq 0,\text{Trace}(S)=1} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j}\langle M_{i,j}M_{i,j}^\top, S\rangle \right\}$$

Smaller rank of optimal solution yields better performance

Theorem. There exists an optimal solution to (SDR) with rank $k \triangleq 1 + \left\lfloor \sqrt{2n + \frac{9}{4}} - \frac{3}{2} \right\rfloor$.

- $$(\text{KMS}) \le (\text{SDR}) \le \max_{S\succeq 0,\text{Trace}(S)=1,\text{rank}(S)=k} \left\{ \min_{\pi\in\Gamma_n} \sum_{i,j=1}^{n} \pi_{i,j}\langle M_{i,j}M_{i,j}^\top, S\rangle \right\}$$

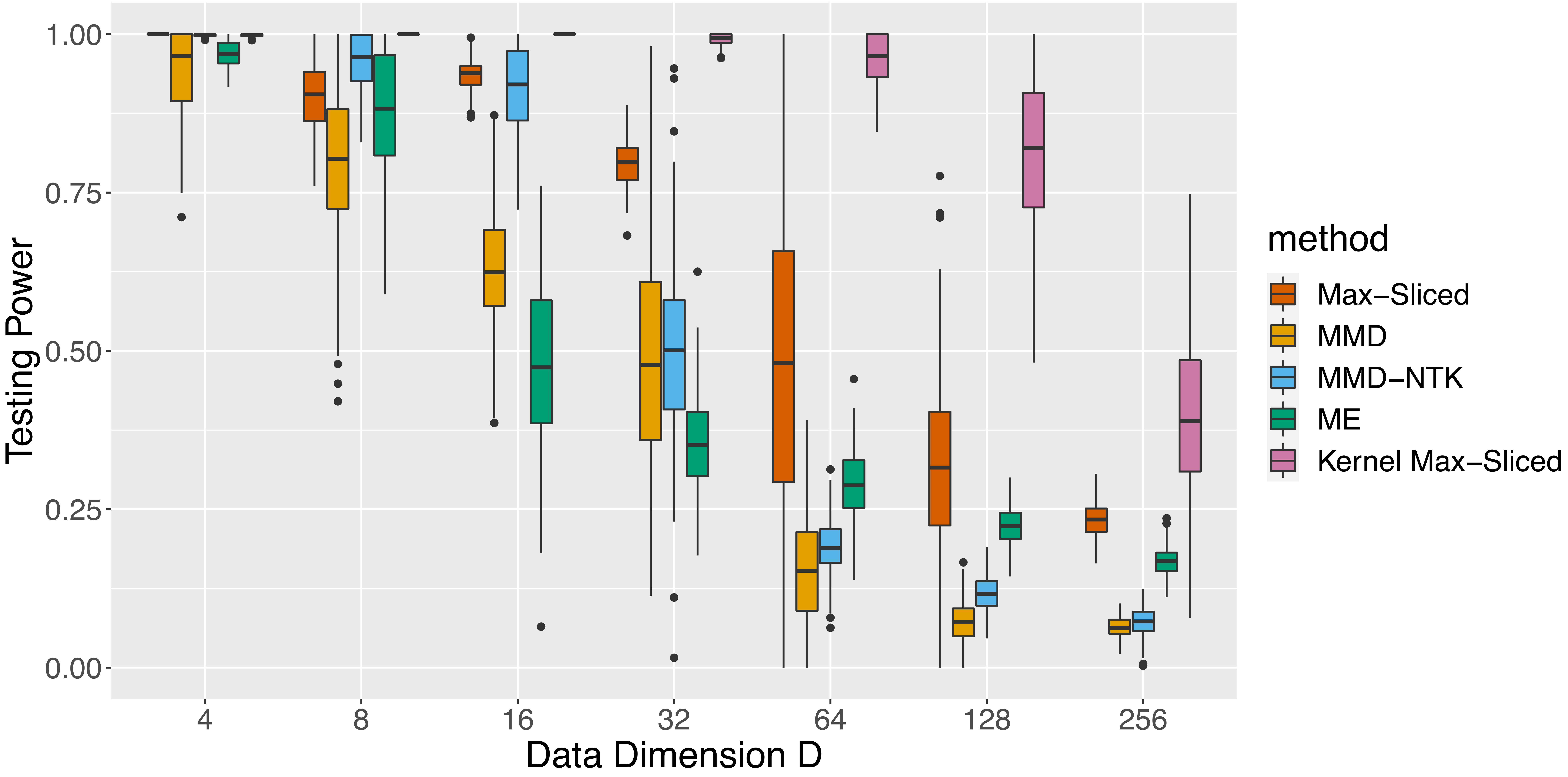


3. Applications and Conclusion

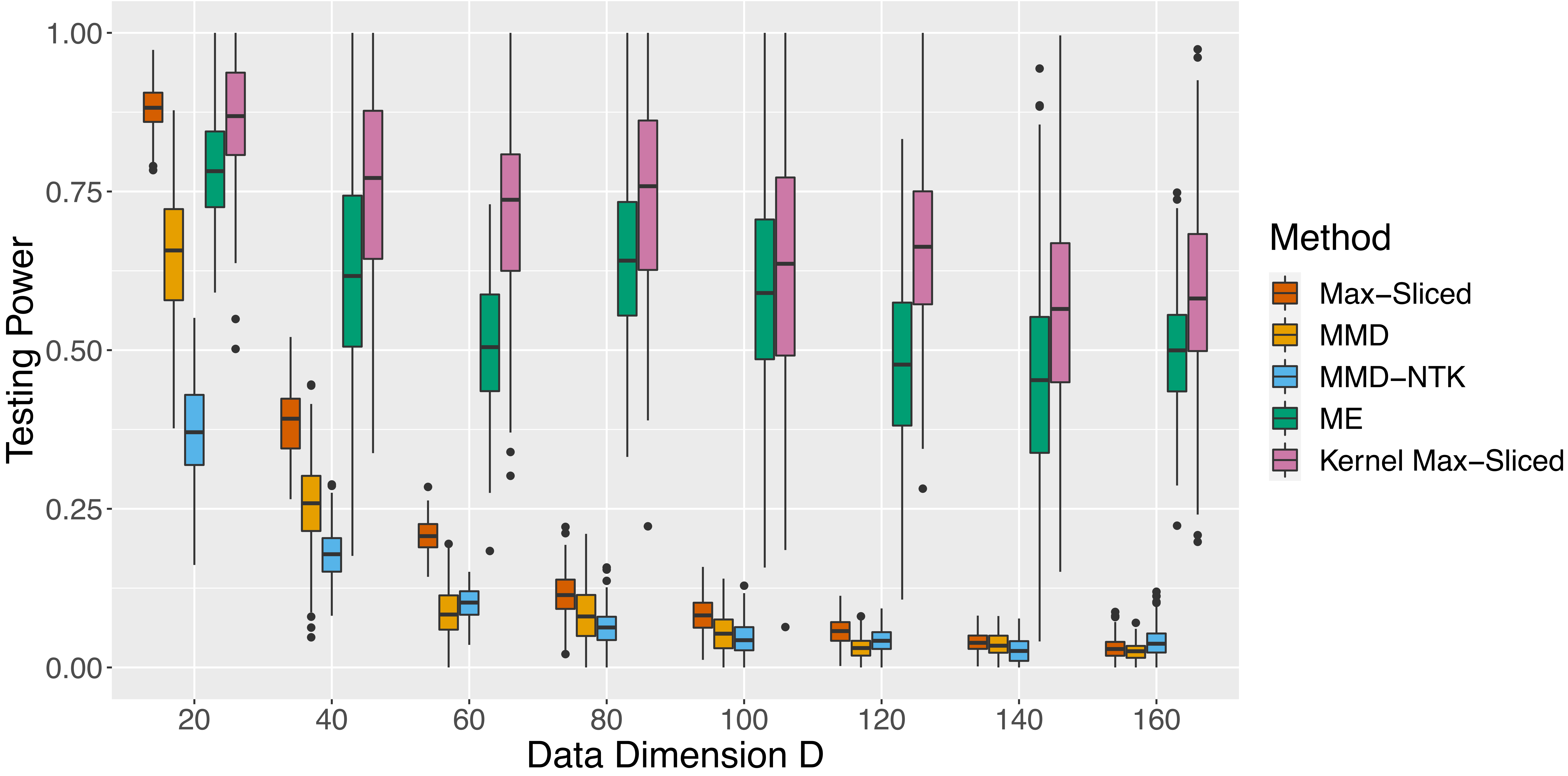
Numerical Experiment Setup

Methodology in Literature	Remarks
Max-Sliced Wasserstein Distance Test **[Deshpande I et al, 2019], [Wang J, 2021]**	Find linear subspace to separate data (**bounded support, log-concave data distribution**)
Maximum Mean Discrepancy (MMD) Test with Optimized Kernel **[Gretton et al. 2012]**	Powerful non-parametric test
MMD Test with Neural Tangent Kernel **[Cheng and Xie, 2021]**	Computationally Efficient with Neural Networks
Mean Embedding Test (ME) **[Jitkrittum et al. 2016]**	Powerful non-parametric test

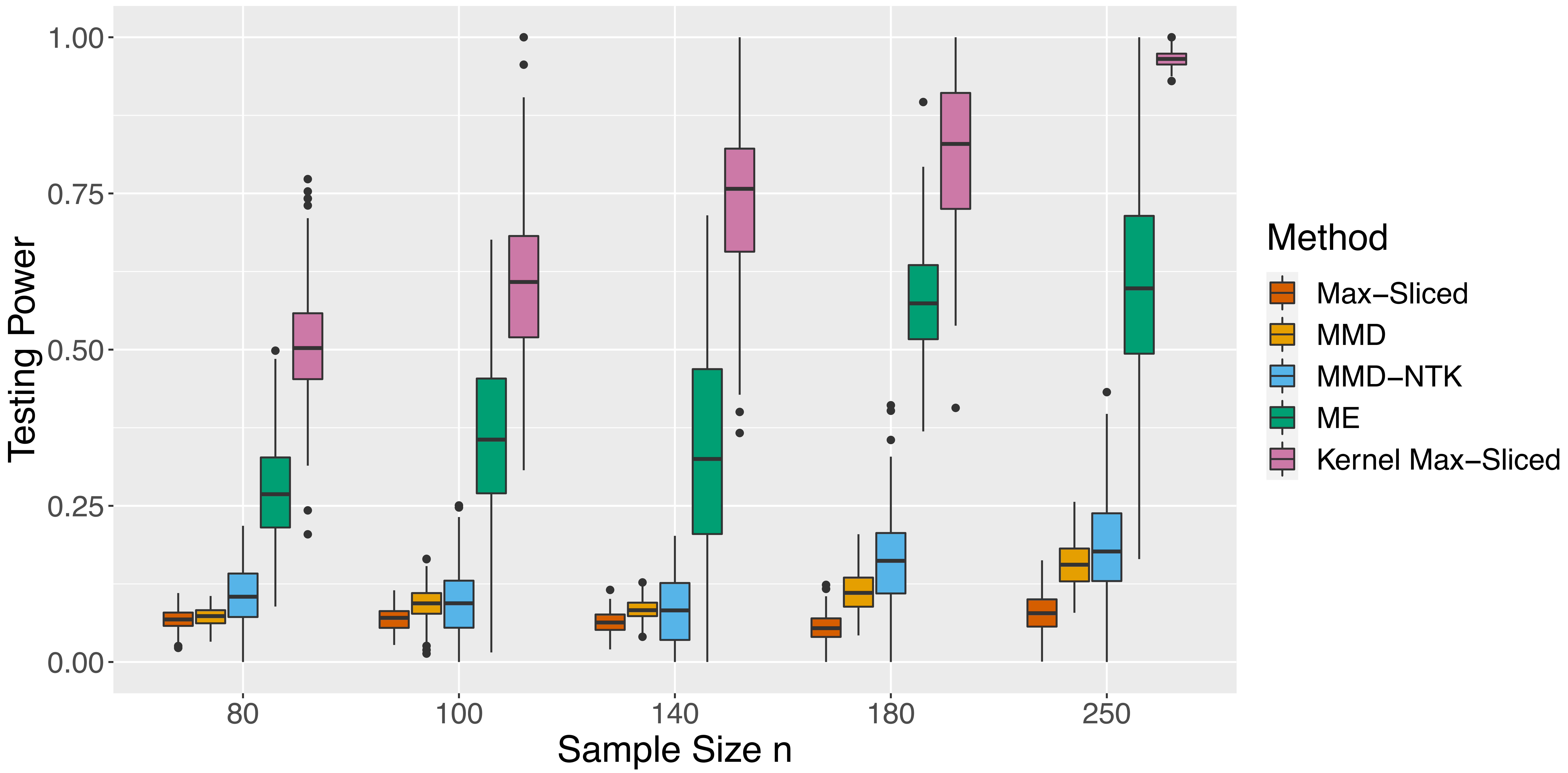
Two-Sample Testing for Gaussian



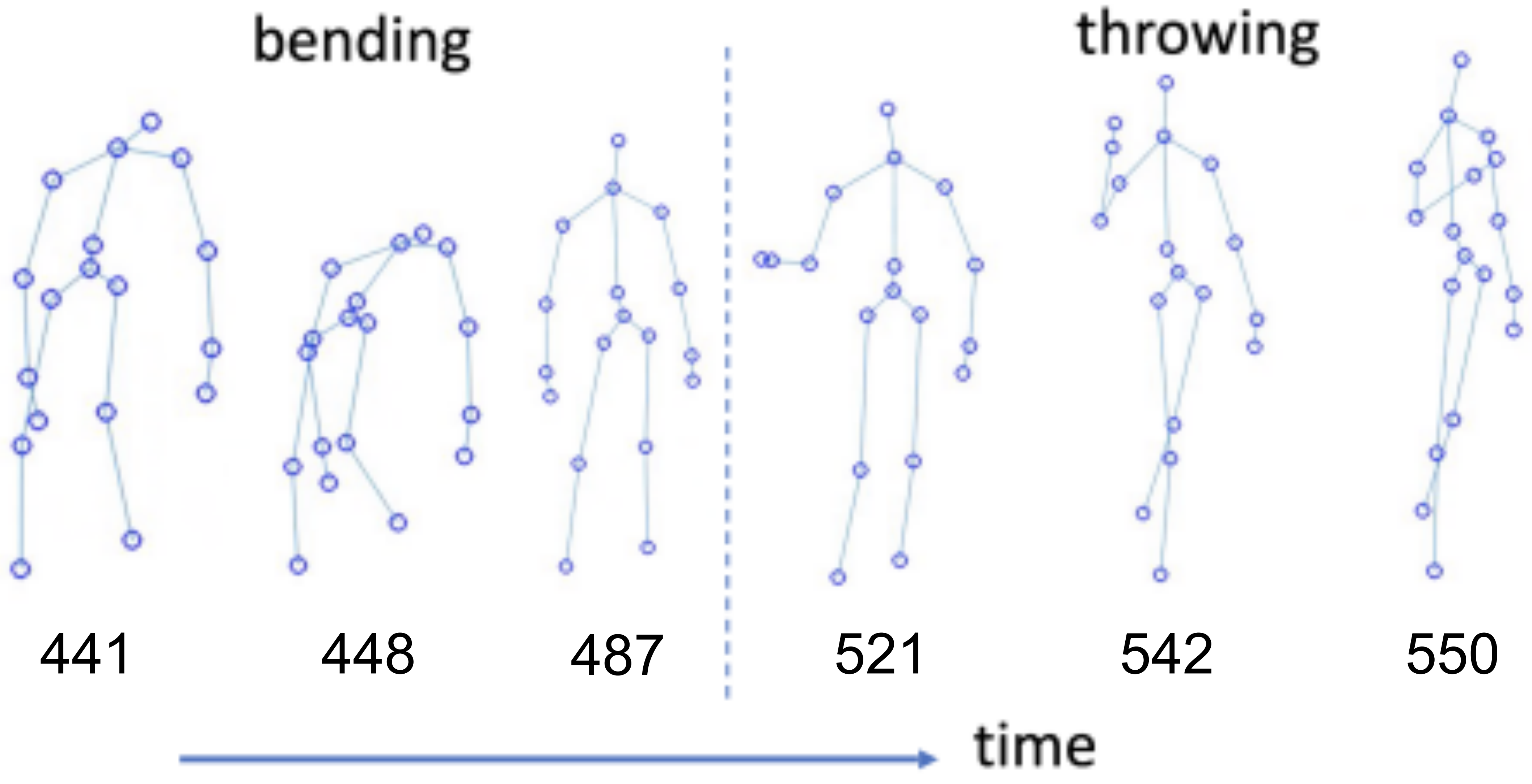
Two-Sample Testing for Gaussian Mixture



Two-Sample Testing for Gaussian Mixture



Online Human Activity Detection

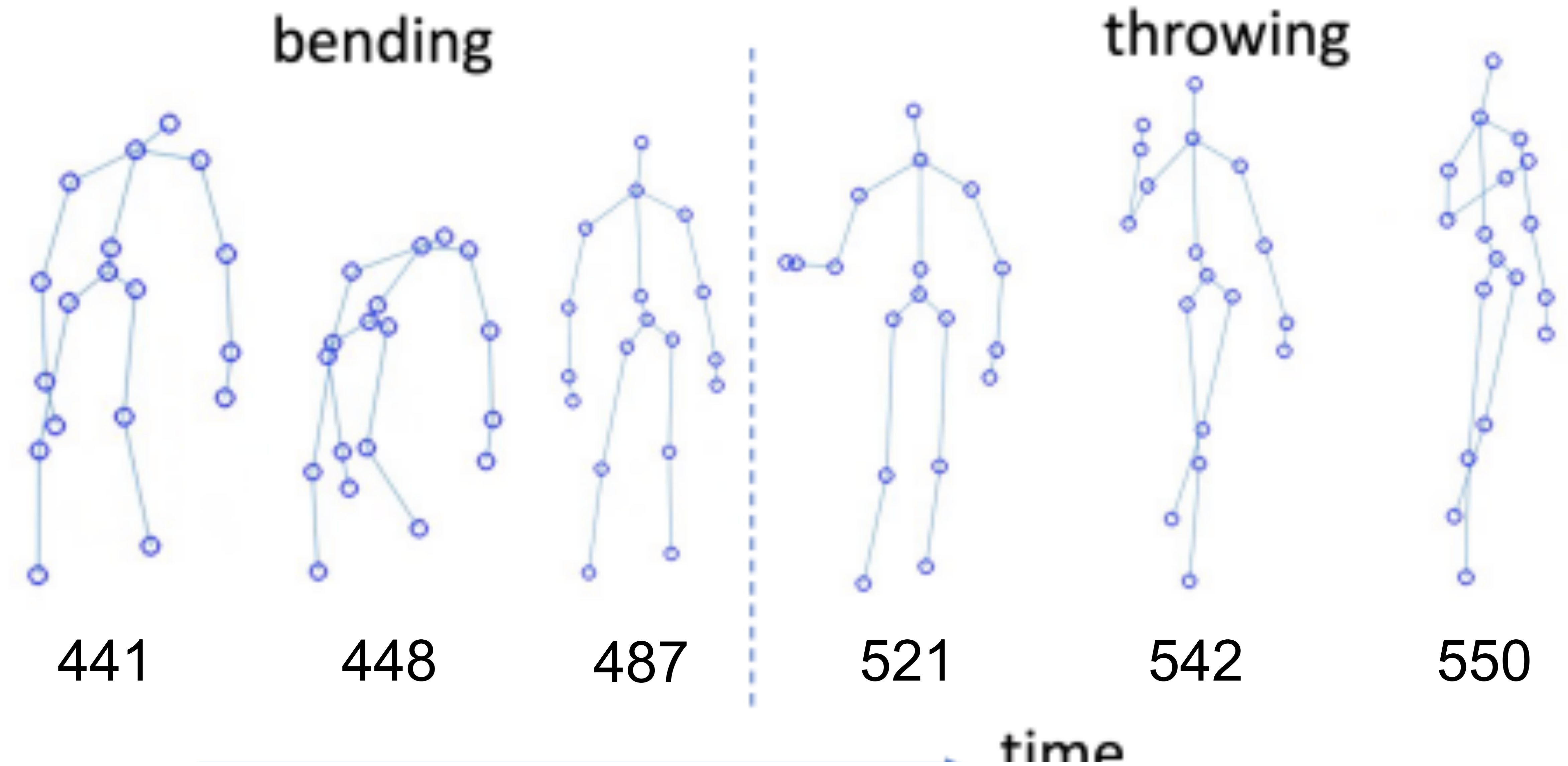


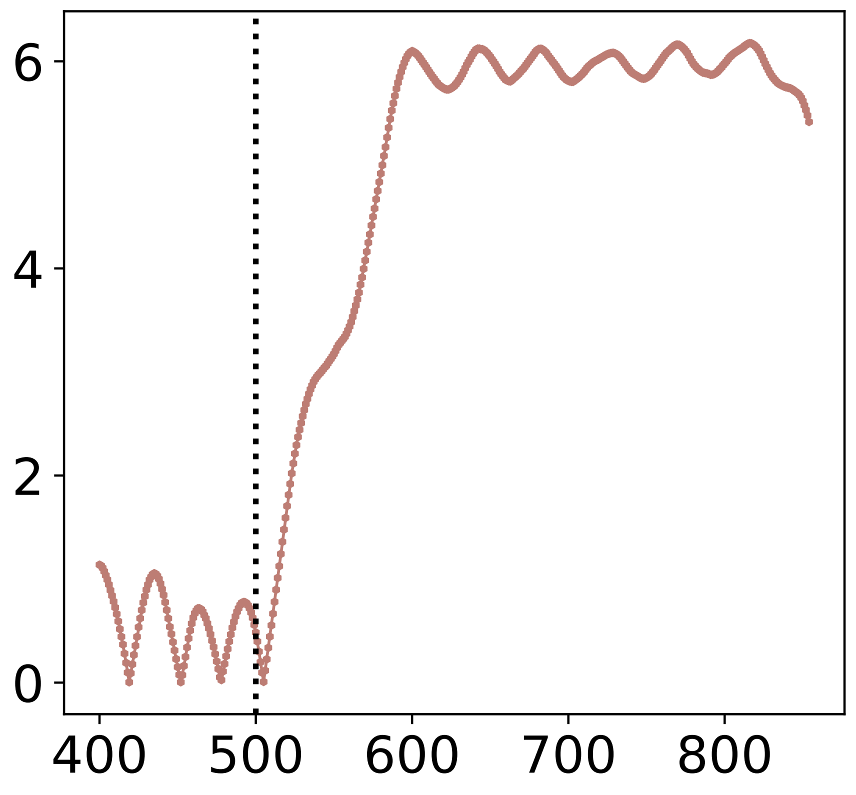
Detection delay using CUSUM

(the smaller the better)

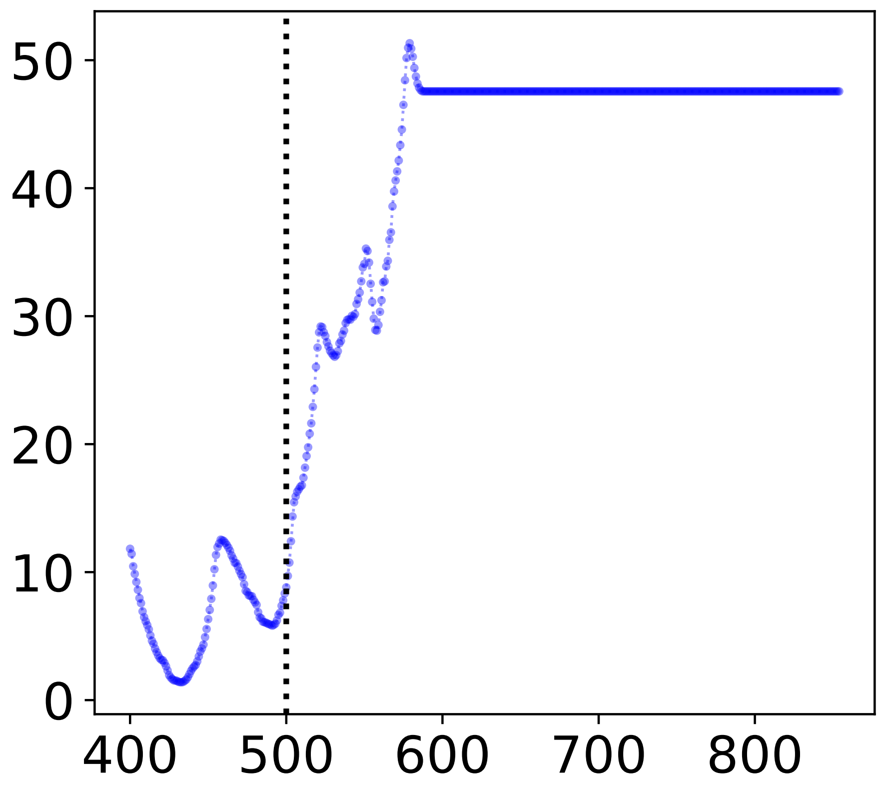
User	Max-Sliced	MMD	MMD-NTK	ME	Kernel Max-Sliced
1	47	73	36	82	**33**
2	9	7	8	97	**1**
3	22	13	15	27	**4**
4	16	83	22	69	**12**
Avg.	23.5	44.0	20.25	68.75	**12.5**

Online Human Activity Detection

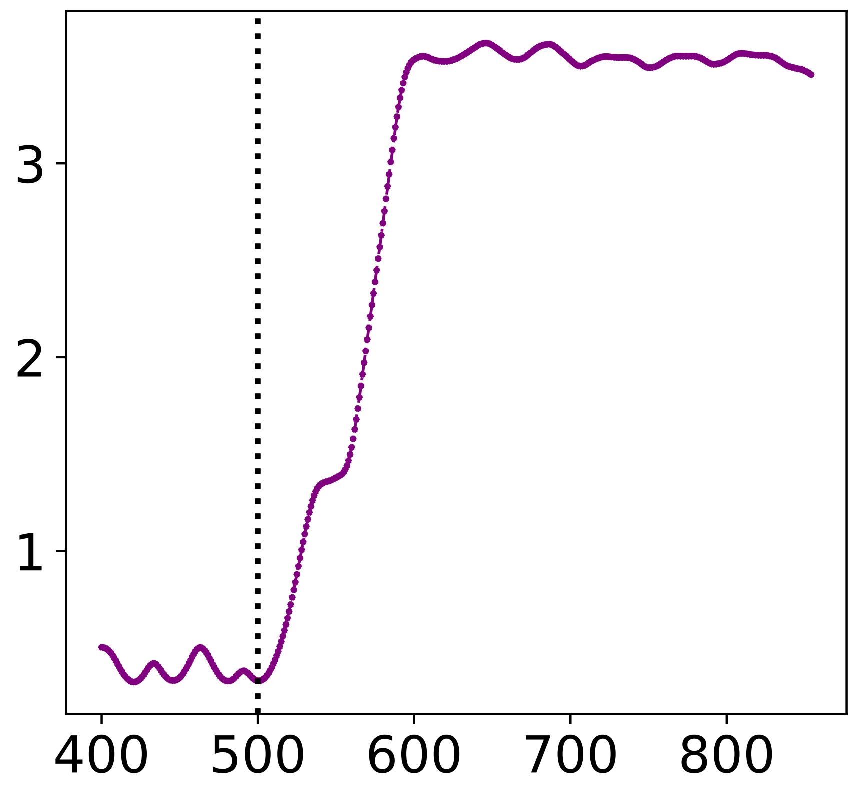




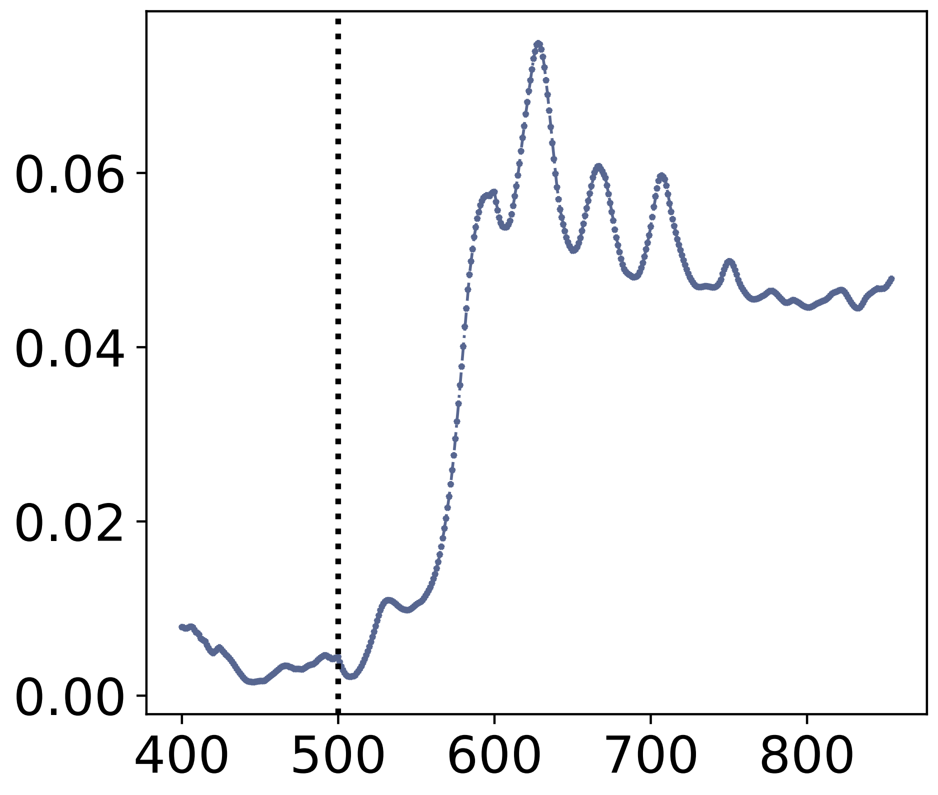
Max-Sliced Wasserstein



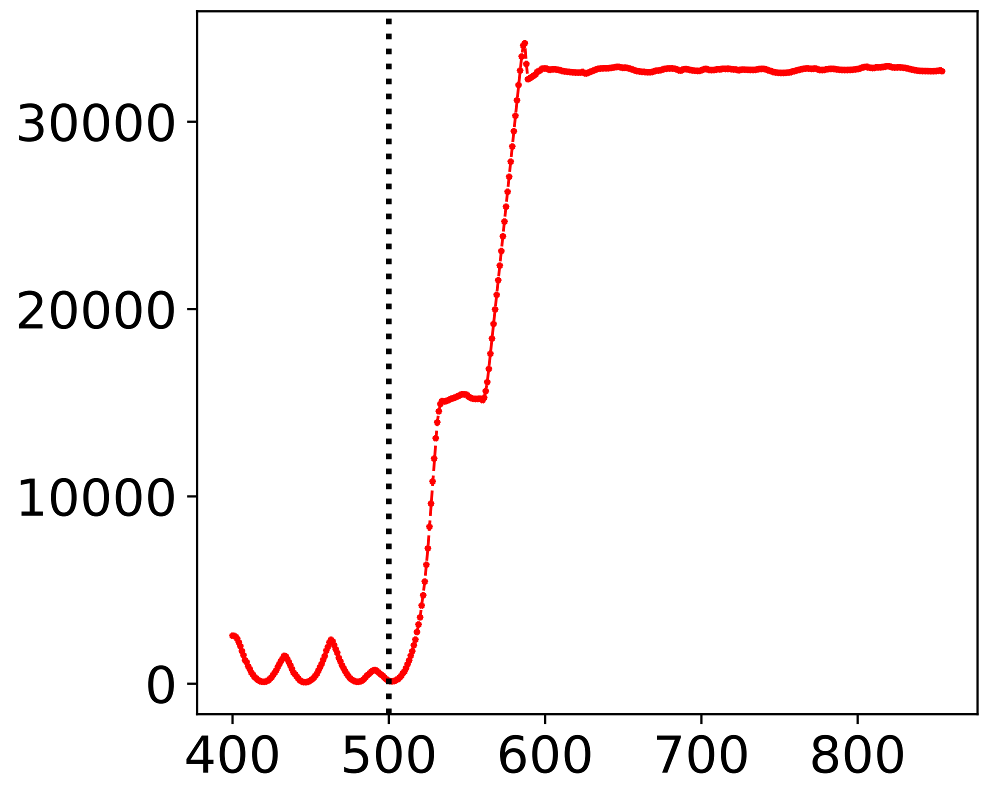
Mean-Embedding



Maximum Mean Discrepancy (MMD)



MMD with Neural Networks



Kernel Max-Sliced Wasserstein

Summary

- A novel **non-parametric** metric for comparing high-dimensional distributions
- **Sharp** finite-sample guarantees
- Computational Guarantees:

  A. Non-concave quadratic maximization problem: $\mathcal{NP}$**-hard**

  B. Approximation algorithm **with performance guarantees**:

- Practical Applications:

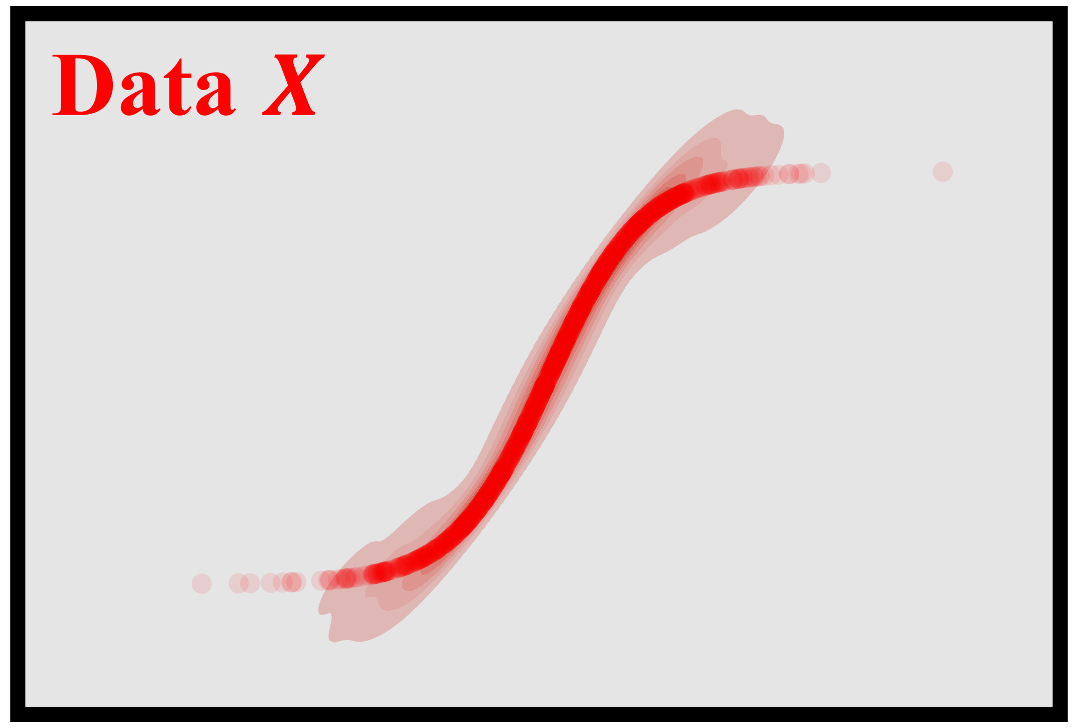
  1. High-dimensional Two-Sample Testing

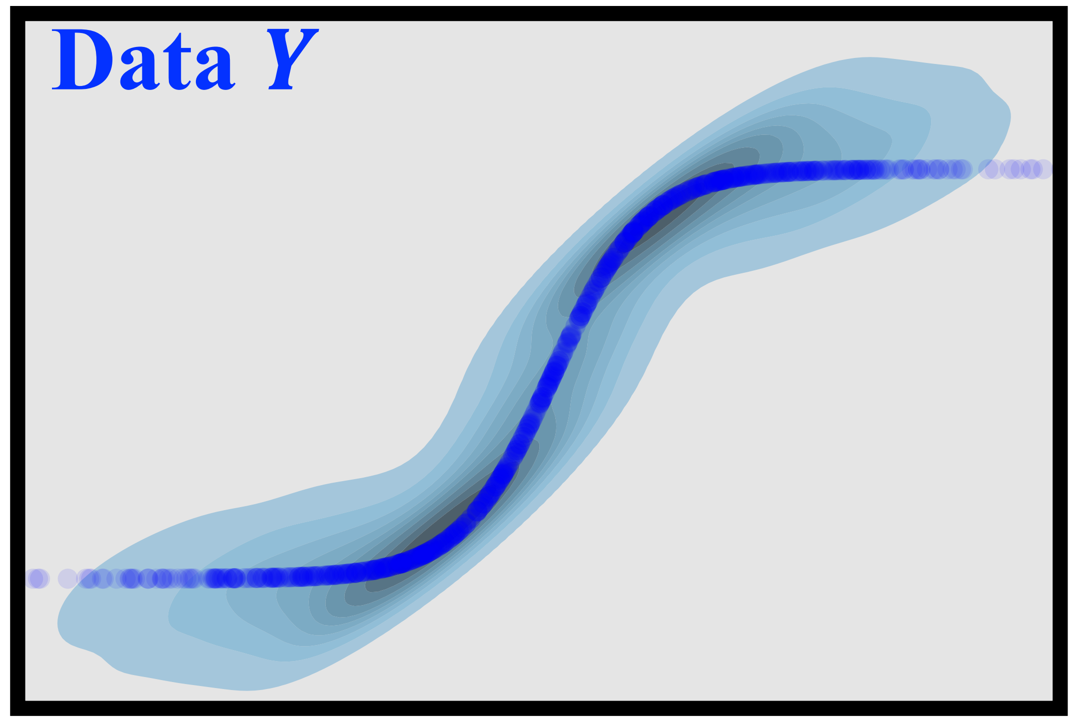
  2. Change-Point Detection

Questions?

$$\mathbf{KMS}(P, Q) = \max_{f \in \mathscr{F}} \mathbf{W}(f_{\#}P, f_{\#}Q)$$

Wang, J., March B., and Yao X., Statistical and Computational Guarantees of Kernel Max-Sliced Wasserstein Distances. arXiv preprint arXiv:2405.15441 (2024).





Nonlinear projector computed from **KMS**