

# **Interpretable Healthcare Prediction**

Jie Wang

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

Date: December 4, 2023

# Table of Contents

- Introduction and Motivation
- Methodology
- Results

# Table of Contents

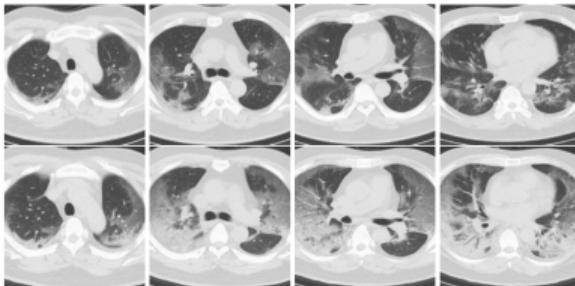
- Introduction and Motivation
- Methodology
- Results

# Variable Selection for Healthcare Datasets

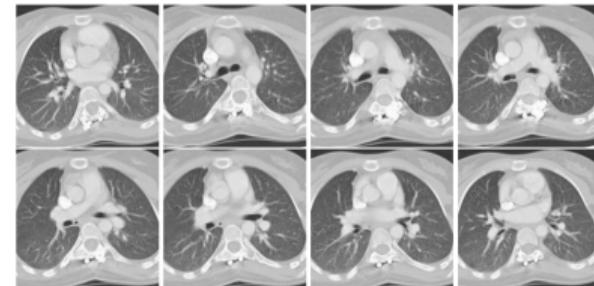
- Two sets of healthcare samples  $\{x_1, \dots, x_n\}$  and  $\{y_1, \dots, y_n\}$  in  $D$  dimensions.
- Q1 :Same distribution or not?

$$\mathcal{H}_0 : P = Q, \quad \mathcal{H}_1 : P \neq Q.$$

- Q2 :Which subset of variables differentiates two sets of samples?



$$\{x_1, \dots, x_n\} \sim P$$



$$\{y_1, \dots, y_n\} \sim Q$$

## Two-Sample Testing

- Goal: given  $X \sim P$  and  $Y \sim Q$ , to determine

$$\mathcal{H}_0 : P = Q, \quad \mathcal{H}_1 : P \neq Q.$$

- Specify a test statistic  $T(X, Y)$  and a threshold  $\tau$ .
- Determine  $\mathcal{H}_1 : P \neq Q$  if  $T(X, Y) > t$ .
- **Kernel two-sample test:** construct  $T(X, Y)$  using MMD (Maximum Mean Discrepancy).
  - Commonly used non-parametric test statistics (Gretton et al. 2021);
  - Apply to high-dimensional and complex data.

# Variable Selection

- Existing variable selection are designed for supervised learning problems:

$$\min \left\{ \frac{1}{n} \sum_{i=1}^n \|y_i - x_i^\top \beta\|_2^2 : \quad \|\beta\|_0 \leq d \right\}$$

$$\min \left\{ \frac{1}{n} \sum_{i=1}^n y_i \beta^\top x_i - \log(1 + \exp(\beta^\top x_i)) : \quad \|\beta\|_0 \leq d \right\}$$

- Solved as mixed integer program (MIP) (Bertsimas, King, Mazumder 2016).
- Convex relaxation formulations, e.g., lasso.
- Our problem is different.

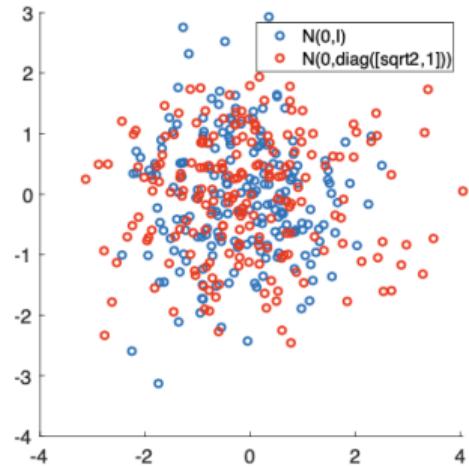
# Why Sparse Logistic Regression is non-applicable?

- Consider an example  $p = \mathcal{N}(0, I_D)$ ,  $q = \mathcal{N}(0, \Sigma)$
- Means  $x_i \sim p$  and  $y_i \sim q$ , distribution only differs in first coordinate, in second-order moment
- Logistic regression corresponds to linear test statistic:

$$T := \frac{1}{n_1} \sum_{i=1}^{n_1} \beta^T x_i - \frac{1}{n_2} \sum_{j=1}^{n_2} \beta^T y_j$$

- $E[T] = 0$  for any  $\beta$
- Logistic regression cannot identify difference in the second-order moments!

$$\Sigma = \text{diag}\{\sqrt{2}, 1\}$$



# Contributions

- Novel formulation: **variable selection** for **kernel** two-sample test<sup>1</sup>.
- Efficient optimization algorithms.
- Healthcare applications:
  - Sepsis diagnosis;
  - Lung cancer data analysis;
  - Skin cancer classification.

---

<sup>1</sup>Variable selection for kernel two-sample test. Wang, Dey, X. arXiv:2302.07415, 2023.

# Table of Contents

- Introduction and Motivation
- Methodology
- Results

# Maximum Mean Discrepancy

- A kernel  $K$  gives a **reproducing kernel Hilbert space**  $\mathcal{H}_K$ .
- (Squared) maximum mean discrepancy statistic:

$$\begin{aligned}\text{MMD}^2(\mu, \nu; K) &\triangleq \sup_{f \in \mathcal{H}_K, \|f\|_{\mathcal{H}_K} \leq 1} \left\{ \mathbb{E}_{\mu}[f] - \mathbb{E}_{\nu}[f] \right\} \\ &= \mathbb{E}_{\mu}[K(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\nu}[K(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}_{\mu, \nu}[K(\mathbf{X}, \mathbf{Y})].\end{aligned}$$

- Empirical MMD estimator:

$$\widehat{\text{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K) \triangleq \frac{1}{n(n-1)} \sum_{i \neq j} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} K(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{n^2} \sum_{i,j} K(\mathbf{x}_i, \mathbf{y}_j).$$

## MMD Variable Selection

Pick the variable selection  $z$  to optimize testing power:

$$\max_{z \in \mathcal{Z}} \quad \widehat{\text{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K_z) - \lambda \widehat{\sigma}^2(\mathbf{x}^n, \mathbf{y}^n; K_z). \quad (\text{MMD-Opt})$$

- $\mathcal{Z} = \left\{ z \in \mathbb{R}^D : \|z\|_2 = 1, \|z\|_0 \leq d \right\};$
- $K_z$  is a kernel **parameterized** by selection vector  $z$ ;
- $\widehat{\sigma}^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$  is the variance estimator of  $\widehat{\text{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$ .

# Algorithm

- For linear kernel, MMD Optimization reduces to **sparse trust region subproblem**:

$$\max_{z \in \mathcal{Z}} \{ z^\top A z + z^\top t \}. \quad (\text{STRS})$$

Closely related to SPCA formulation:



Search...  
Help | Adv

Statistics > Machine Learning

[Submitted on 28 Aug 2020]

## Exact and Approximation Algorithms for Sparse PCA

Yongchun Li, Weijun Xie

Sparse PCA (SPCA) is a fundamental model in machine learning and data analytics, which has witnessed a variety of application areas such as finance, manufacturing, biology, healthcare. To select a prespecified-size principal submatrix from a covariance matrix to maximize its largest eigenvalue for the better interpretability purpose, SPCA advances the conventional PCA with both feature selection and dimensionality reduction. This paper proposes two exact mixed-integer SDPs (MISDPs) by exploiting the spectral decomposition of the covariance matrix and the properties of the largest eigenvalues. We then analyze the theoretical optimality gaps of their continuous relaxation values and prove that they are stronger than that of the state-of-art one. We further show that the continuous relaxations of two MISDPs can be recast as saddle point problems without involving semi-definite cones, and thus can be effectively solved by first-order methods such as the subgradient method. Since off-the-shelf solvers, in general, have difficulty in solving MISDPs, we approximate SPCA with arbitrary accuracy by a mixed-integer linear program (MILP) of a similar size as MISDPs. To be more scalable, we also analyze greedy and local search algorithms, prove their first-known approximation ratios, and show that the approximation ratios are tight. Our numerical study demonstrates that the continuous relaxation values of the proposed MISDPs are quite close to optimality, the proposed MILP model can solve small and medium-size instances to optimality, and the approximation algorithms work very well for all the instances. Finally, we extend the analyses to Rank-one Sparse SVD (R1-SSVD) with non-symmetric matrices and Sparse Fair PCA (SFPCA) when there are multiple covariance matrices, each corresponding to a protected group.

# Algorithm

- For linear kernel, MMD Optimization reduces to **sparse trust region subproblem**:

$$\max_{z \in \mathcal{Z}} \{ z^\top A z + z^\top t \}. \quad (\text{STRS})$$

Problem (STRS) is equivalent to

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

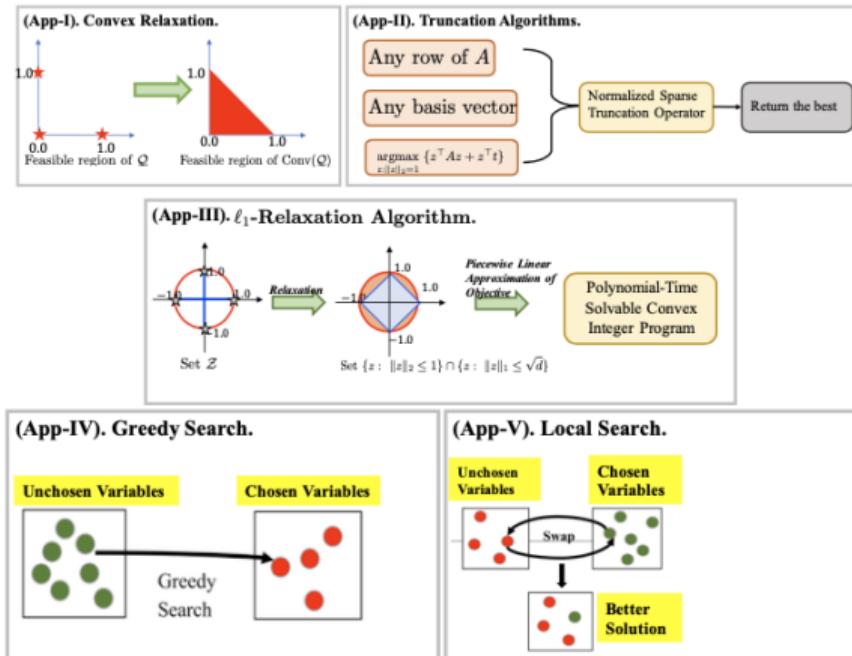
where  $\tilde{A} = [0, t^\top; t, A]$ , and the set

$$\mathcal{Q} = \{q \in \{0, 1\}^D : \sum_{k \in [D]} q_k \leq d\}.$$

# Algorithm

- For linear kernel, MMD Optimization reduces to **sparse trust region subproblem**:

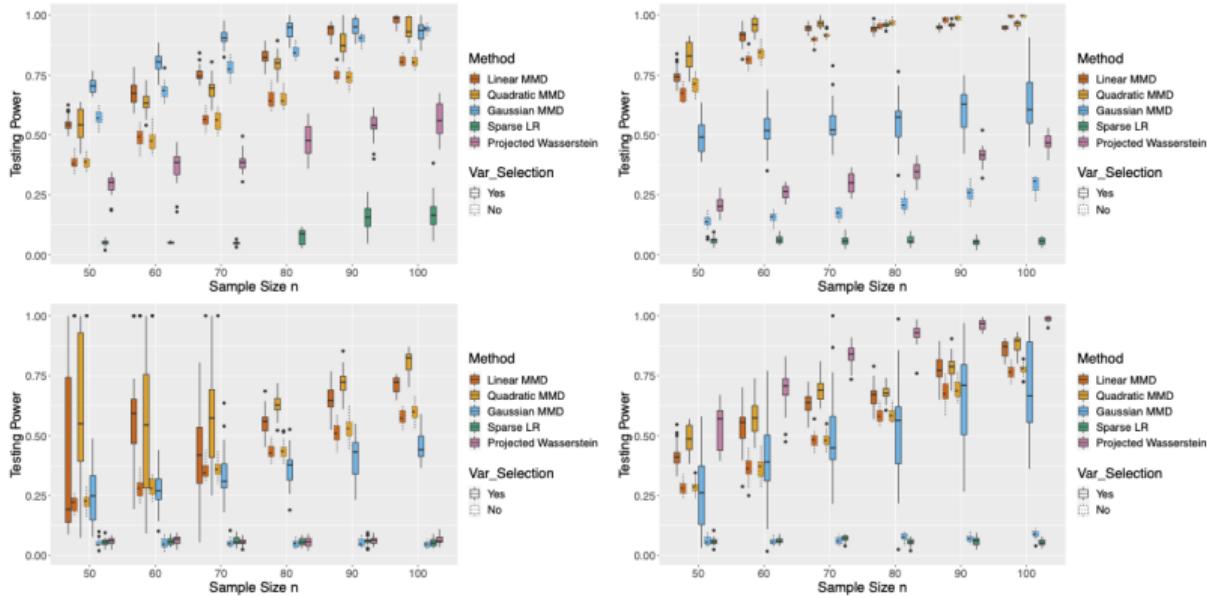
$$\max_{z \in \mathcal{Z}} \{ z^\top A z + z^\top t \}. \quad (\text{STRS})$$



# Table of Contents

- Introduction and Motivation
- Methodology
- Results

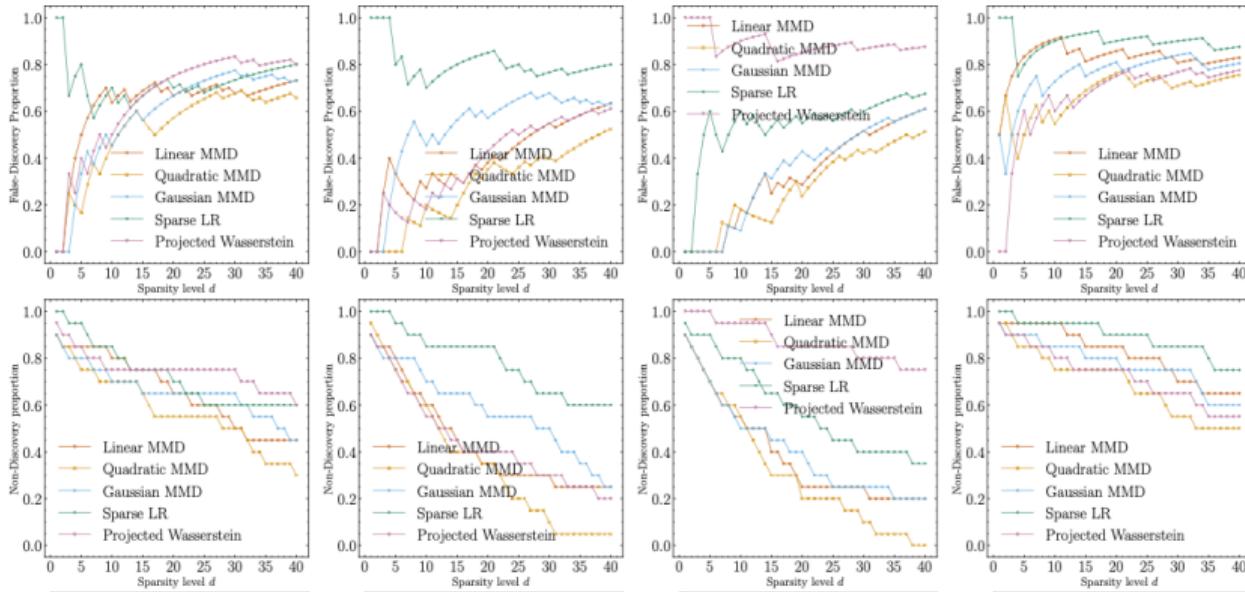
# Synthetic Datasets



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

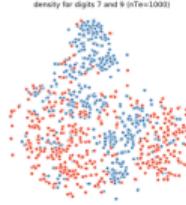
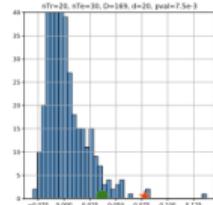
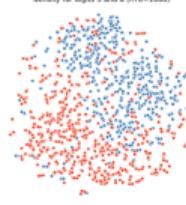
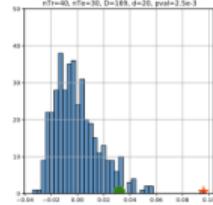
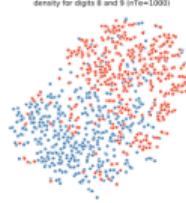
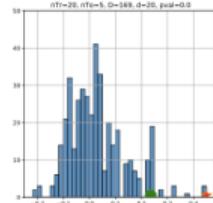
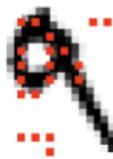
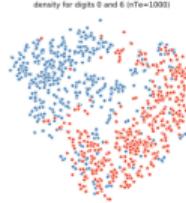
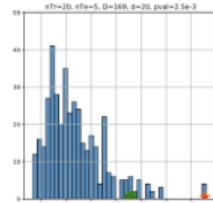
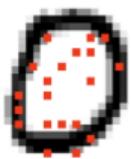
- Algorithm complexity: Linear < **Quadratic** < Gaussian
- Testing Performance: **Quadratic** > Gaussian > Linear

# Synthetic Datasets



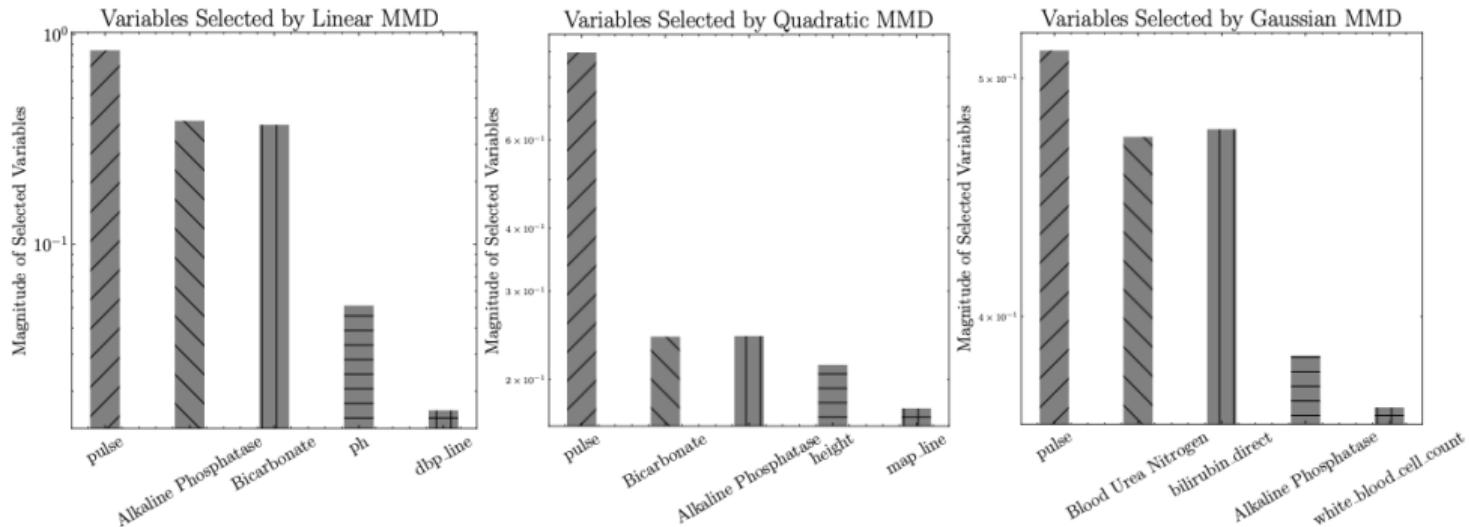
- Algorithm complexity: Linear < **Quadratic** < Gaussian
- Testing Performance: **Quadratic** > Gaussian > Linear

# MNIST Dataset



# Sepsis Prediction

- Medical records from Emory hospital with dimension  $D = 39^2$ ;
- $m = 20771$  healthy people and  $n = 2891$  Sepsis patients;
- Training sample size  $m_{Tr} = 20000, n_{Tr} = 2000$ .



<sup>2</sup>Wang J, Moore R, Xie Y, et al. Improving sepsis prediction model generalization with optimal transport[C]/Machine Learning for Health. PMLR, 2022: 474-488.

## Sepsis Prediction

- Medical records from Emory hospital with dimension  $D = 39^2$ ;
- $m = 20771$  healthy people and  $n = 2891$  Sepsis patients;
- Training sample size  $m_{Tr} = 20000, n_{Tr} = 2000$ .

**Table 2** Averaged testing power for the sepsis prediction.

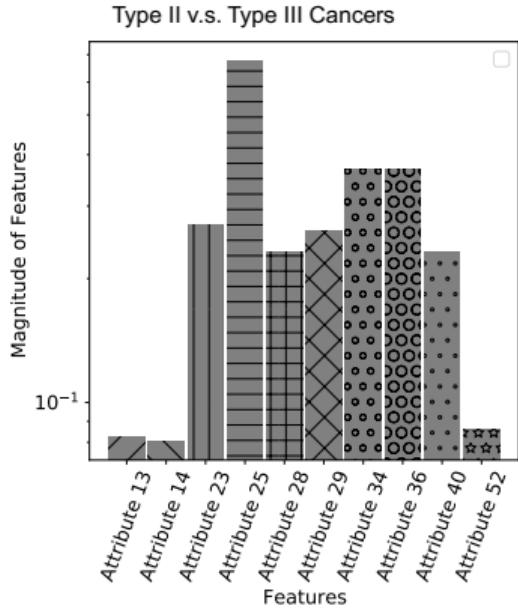
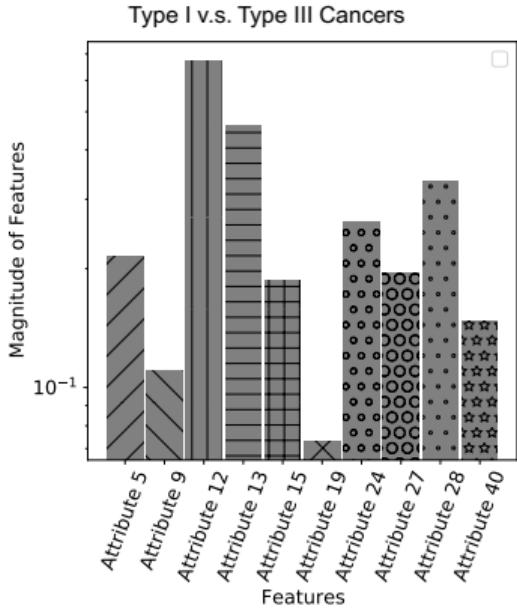
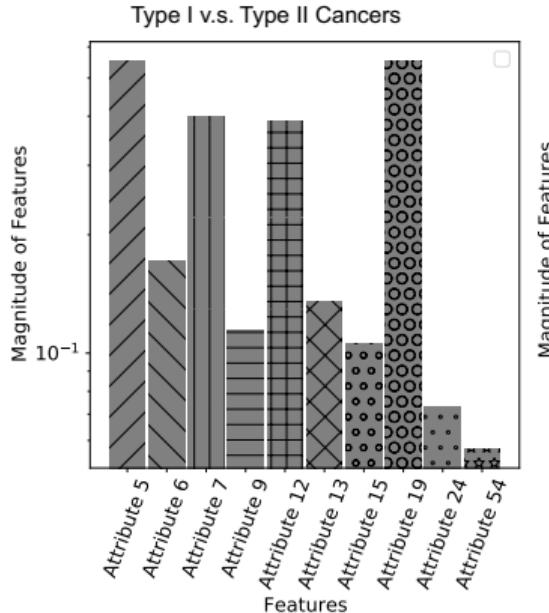
Linear MMD	Quadratic MMD	Gaussian MMD	Logistic Regression	Projected Wasserstein
0.835	0.915	0.784	0.771	0.749

---

<sup>2</sup>Wang J, Moore R, Xie Y, et al. Improving sepsis prediction model generalization with optimal transport[C]/Machine Learning for Health. PMLR, 2022: 474-488.

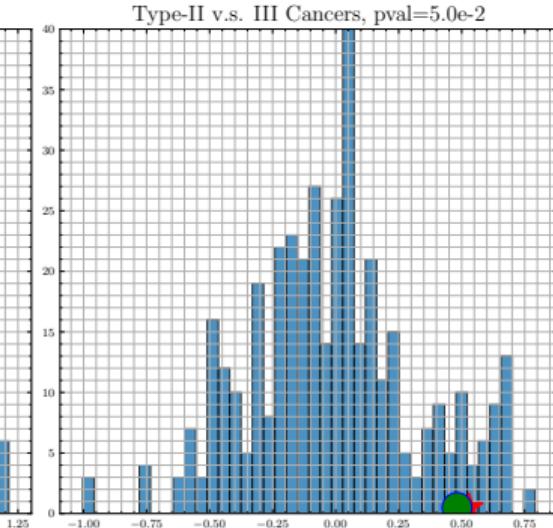
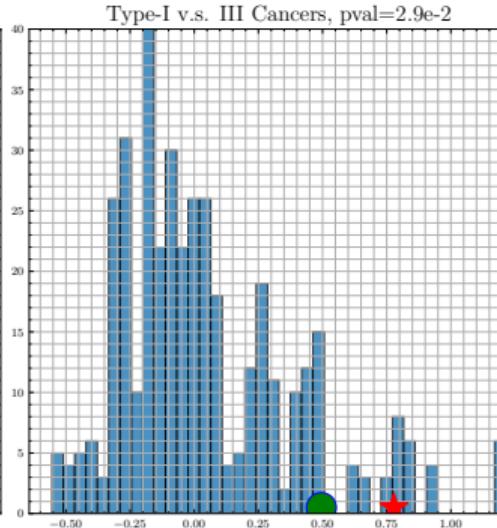
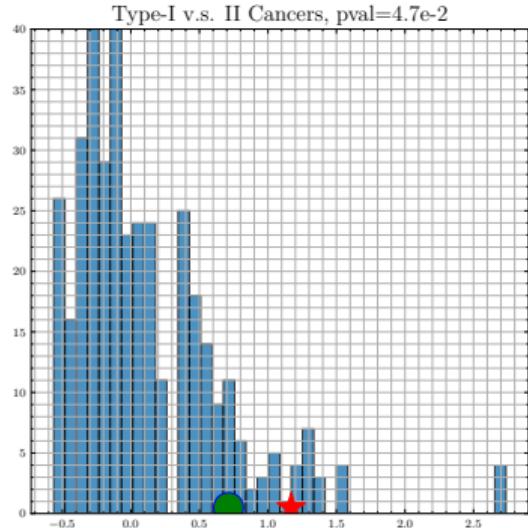
# UCI Lung Cancer Dataset Analysis

- 56 attributes for 32 patients;
- 3 types of pathological lung cancers.

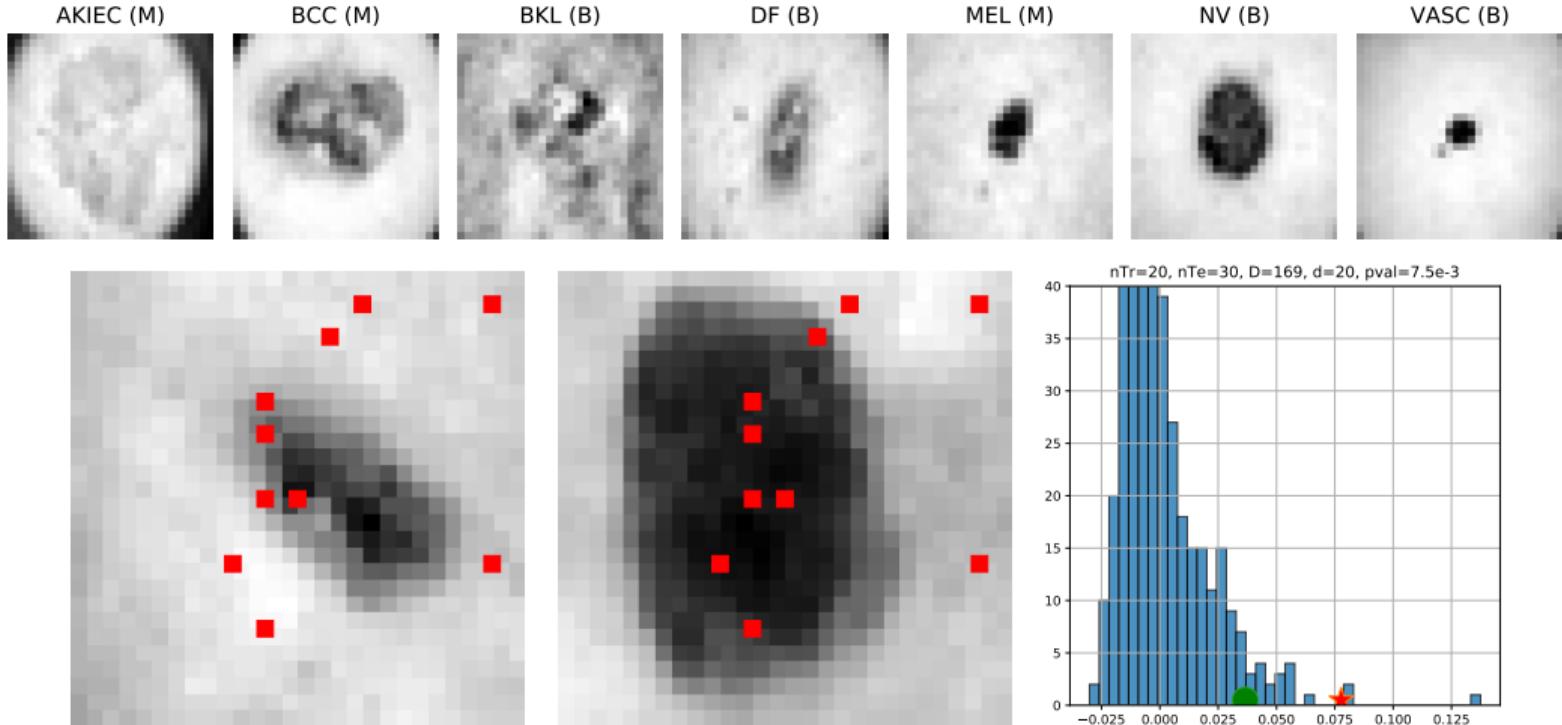


# UCI Lung Cancer Dataset Analysis

- 56 attributes for 32 patients;
- 3 types of pathological lung cancers.



# Skin Cancer Image Dataset Classification



## Discussion

- Flexible variable selection framework for interpretable healthcare data analysis.
- Combination of statistical hypothesis testing and mixed-integer optimization.
- Applications in medical records analysis and image classification.