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

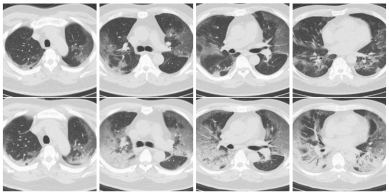
Table of Contents

Variable Selection for Healthcare Datasets

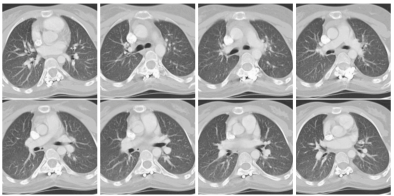
- Two sets of healthcare samples $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ in D dimensions.
- Q1 :Same distribution or not?

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

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



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



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

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 τ.
- 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:

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

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

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

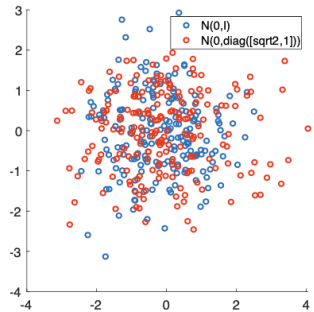
Why Sparse Logistic Regression is non-applicable?

- 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 cooridnate, in second-order moment
- Logistic regression corresponds to linear test statistic:

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

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

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



Contributions

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

[1] Variable selection for kernel two-sample test. Wang, Dey, X. arXiv:2302.07415, 2023.

Table of Contents

Maximum Mean Discrepancy

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

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

- Empirical MMD estimator:

$$\widehat{\mathrm{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K) \triangleq \frac{1}{n(n-1)} \sum_{i \neq j} K(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} K(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} K(x_i, y_j).$$

MMD Variable Selection

Pick the variable selection z to optimize testing power:

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

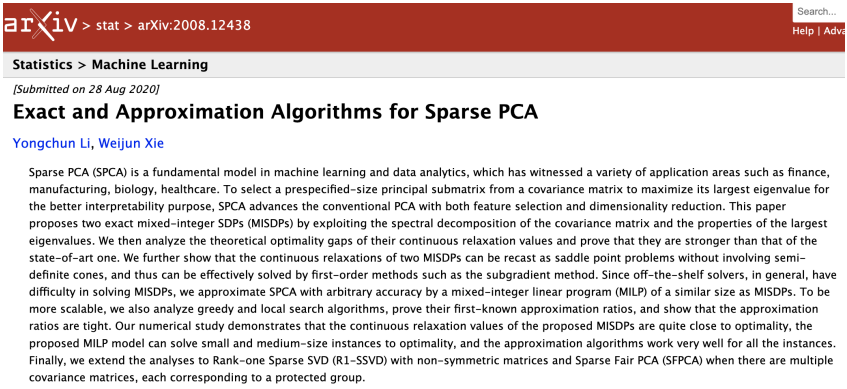
- $\mathcal{Z} = \left\{ z \in \mathbb{R}^D : \ \|z\|_2 = 1, \|z\|_0 \leq d \right\}$;
- K_z is a kernel **parameterized** by selection vector z;
- $\widehat{\sigma}^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$ is the variance estimator of $\widehat{\mathrm{MMD}}^2(\mathbf{x}^n, \mathbf{y}^n; K_z)$.

Algorithm

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

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

Closely related to SPCA formulation:



Algorithm

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

$$\max_{z \in \mathcal{Z}} \ \{z^\top A z + z^\top t\}. \tag{STRS}$$

Problem (STRS) is equivalent to

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

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

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

Algorithm

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

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

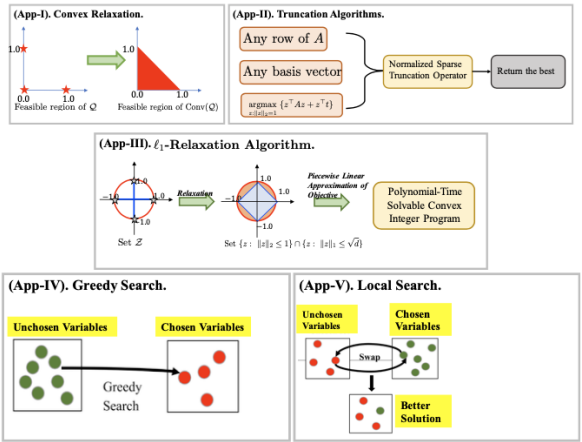


Table of Contents

Synthetic Datasets

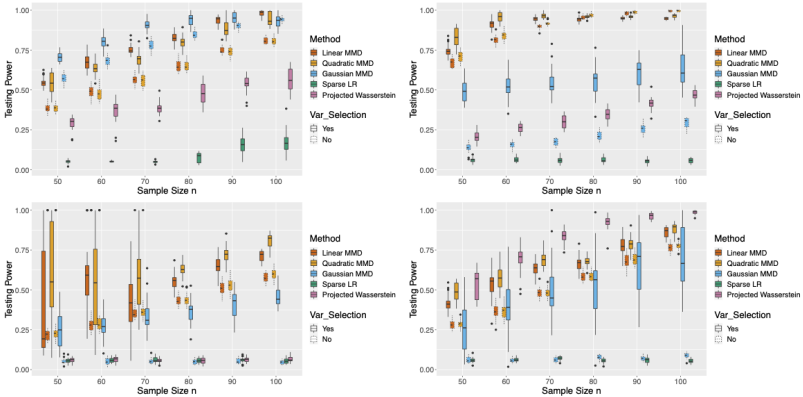
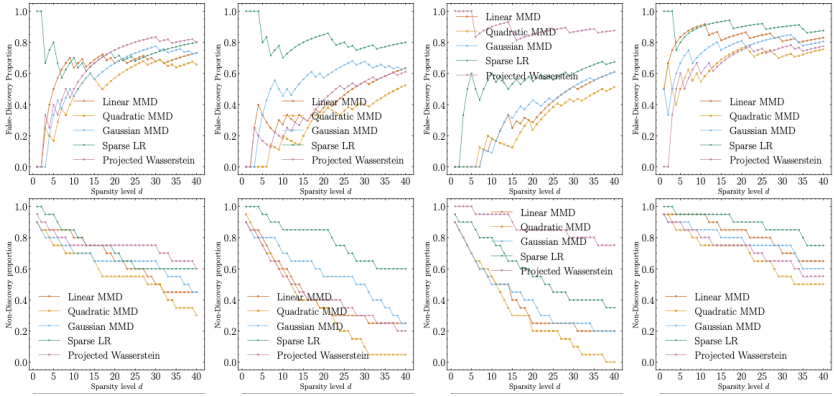


Figure 2 Testing power of various two-sample tests with different choices of sample size n. Here we fix parameters $D = 100,\ d_{\text{true}} = 20,\ d = 20$ and control the type-I error $\alpha_{\text{level}} = 0.05$. Plots from top to bottom correspond to four different types of synthetic datasets.

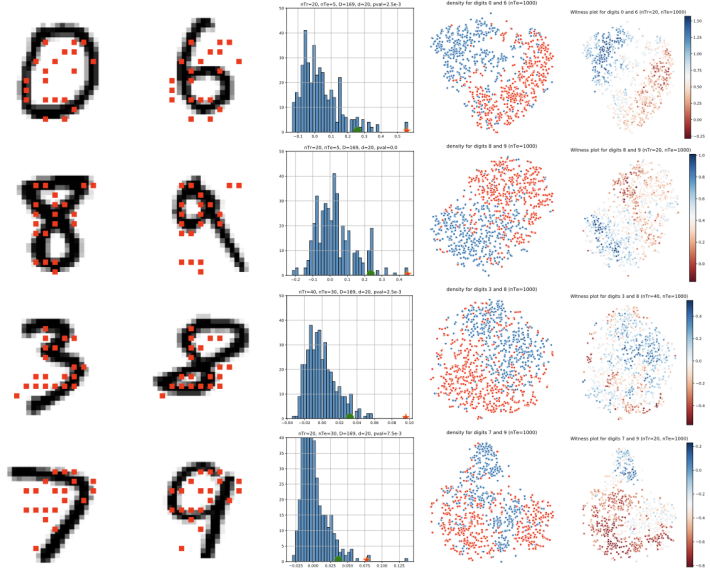
- Algorithm complexity: Linear $<$ Quadratic $<$ Gaussian
- Testing Performance: Quadratic $>$ Gaussian $>$ Linear

Synthetic Datasets



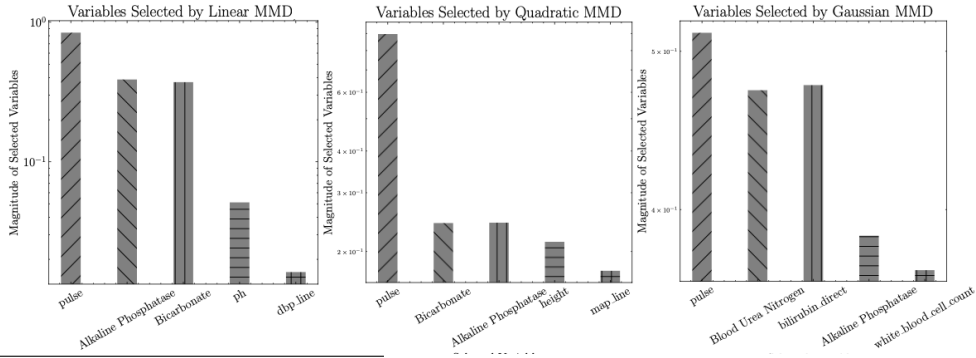
- Algorithm complexity: Linear $<$ Quadratic $<$ Gaussian
- Testing Performance: Quadratic $>$ Gaussian $>$ Linear

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 $m_{\mathsf{Tr}} = 20000, n_{\mathsf{Tr}} = 2000$.



[2]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 $m_{\mathsf{Tr}} = 20000, n_{\mathsf{Tr}} = 2000$.

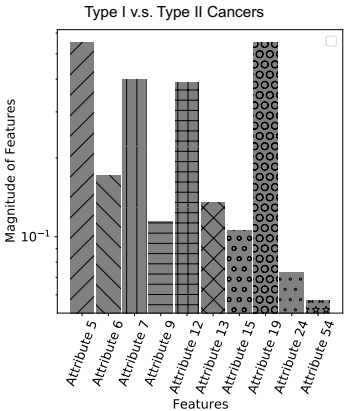
Table 2 Averaged testing power for the sepsis prediction.

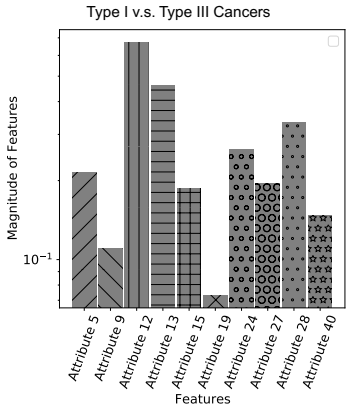
Linear MMD	Quadratic MMD	Gaussian MMD	Logistic Regression	Projected Wasserstein
0.835	0.915	0.784	0.771	0.749

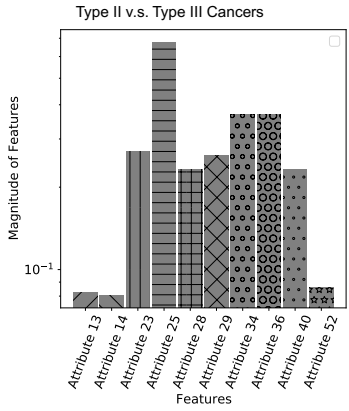
[2]Wang J, Moore R, Xie Y, et al. Improving sepsis prediction model generalization with optimal transport[C]/Machine Learning for Health. PMLR, 2022: 474-488.

UCI Lung Cancer Dataset Analysis

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

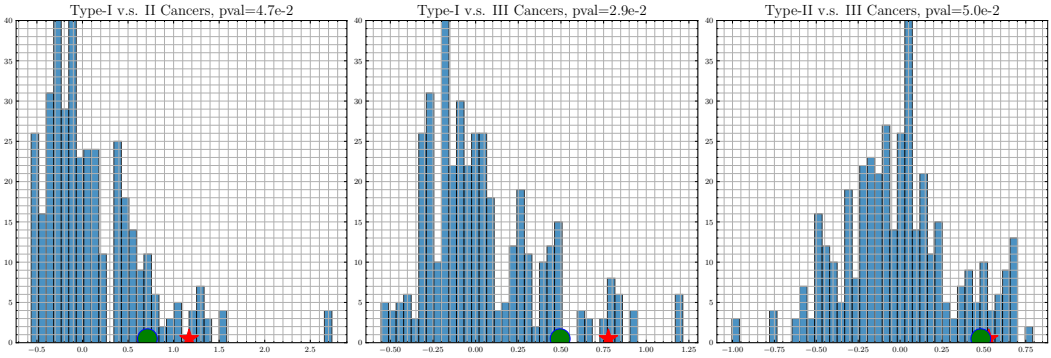




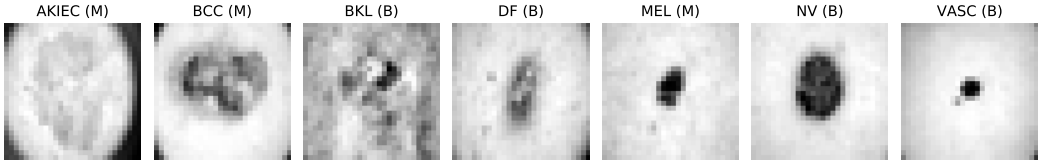


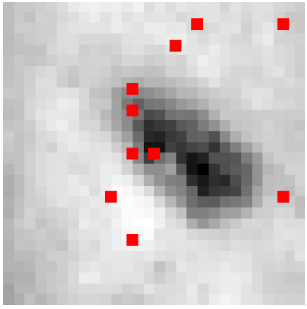
UCI Lung Cancer Dataset Analysis

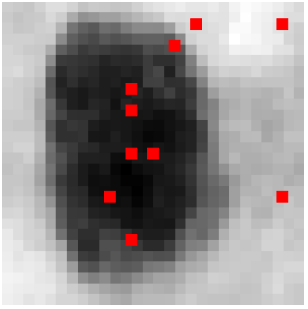
- 56 attributes for 32 patients;
- 3 types of pathological lung cancers.

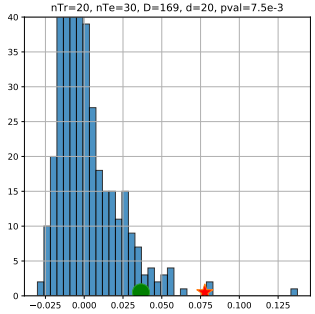


Skin Cancer Image Dataset Classification









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.