Distribution Regression with Sliced Wasserstein Kernels

Dimitri Meunier[1], Massimiliano Pontil[2,3], Carlo Ciliberto[3]

June 27, 2022

Spotlight 39th International Conference on Machine Learning (ICML 2022)

[1] Gatsby Computational Neuroscience Unit, UCL, London; [2]Italian Institute of Technology, Genoa;
[3]Department of Computer Science, UCL, London

Motivations

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{`healthy'}, \text{`diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{'healthy'}, \text{'diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc
- Distribution of health indicators $\mathbb{P}_i \in \mathcal{P}(\mathbb{R}^r)$ $(1 \leq i \leq m)$

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{'healthy'}, \text{'diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc
- Distribution of health indicators $\mathbb{P}_i \in \mathcal{P}(\mathbb{R}^r)$ $(1 \leq i \leq m)$
- Repeated measurements over time $X_{i,t} \sim \mathbb{P}_i$ $(1 \leq t \leq n_i)$

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{'healthy'}, \text{'diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc
- Distribution of health indicators $\mathbb{P}_i \in \mathcal{P}(\mathbb{R}^r)$ $(1 \leq i \leq m)$
- Repeated measurements over time $X_{i,t} \sim \mathbb{P}_i$ $(1 \leq t \leq n_i)$
  $\mathcal{X}_i := \{X_{i,1}, \ldots, X_{i,n_i}\} \in (\mathbb{R}^r)^{n_i}$, $\mathcal{D} := \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{'healthy'}, \text{'diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc
- Distribution of health indicators $\mathbb{P}_i \in \mathcal{P}(\mathbb{R}^r)$ $(1 \leq i \leq m)$
- Repeated measurements over time $X_{i,t} \sim \mathbb{P}_i$ $(1 \leq t \leq n_i)$
  $\mathcal{X}_i := \{X_{i,1}, \ldots, X_{i,n_i}\} \in (\mathbb{R}^r)^{n_i}$, $\mathcal{D} := \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$
- **Standard (Euclidean) Regression:** Aggregate features $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_j$, lose information!

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{'healthy'}, \text{'diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc
- Distribution of health indicators $\mathbb{P}_i \in \mathcal{P}(\mathbb{R}^r)$ $(1 \leq i \leq m)$
- Repeated measurements over time $X_{i,t} \sim \mathbb{P}_i$ $(1 \leq t \leq n_i)$
  $\mathcal{X}_i := \{X_{i,1}, \dots, X_{i,n_i}\} \in (\mathbb{R}^r)^{n_i}$, $\mathcal{D} := \{\mathcal{X}_1, \dots, \mathcal{X}_m\}$
- ~~**Standard (Euclidean) Regression:** Aggregate features $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j$, lose information!~~
- **Distribution Regression:** Learn directly from
  $\hat{\mathbb{P}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{i,j}} \approx \mathbb{P}_i$

Motivations: patient diagnosis

Diagnosing patients, $Y = \{\text{'healthy'}, \text{'diseased'}\}$

- m patients with access to r health indicators: heart rate, blood pressure, chemical concentrations in blood etc
- Distribution of health indicators $\mathbb{P}_i \in \mathcal{P}(\mathbb{R}^r)$ $(1 \leq i \leq m)$
- Repeated measurements over time $X_{i,t} \sim \mathbb{P}_i$ $(1 \leq t \leq n_i)$ $\mathcal{X}_i := \{X_{i,1}, \ldots, X_{i,n_i}\} \in (\mathbb{R}^r)^{n_i}$, $\mathcal{D} := \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$
- ~~**Standard (Euclidean) Regression:** Aggregate features $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j$, lose information!~~
- **Distribution Regression:** Learn directly from $\hat{\mathbb{P}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{i,j}} \approx \mathbb{P}_i$

Goal:

$$\hat{f}_{\mathcal{D}} : \mathcal{P}(\mathbb{R}^r) \longrightarrow \{\text{'healthy'}, \text{'diseased'}\}$$

$$\hat{\mathbb{P}}_{\text{new patients}} \longmapsto y_{\text{new patients}}$$

Kernel Regression versus Distribution Regression

Kernel Regression versus Distribution Regression

1. **Standard Regression:** predict real-valued response Y from a vector-valued covariate $X \in \mathbb{R}^r$

Kernel Regression versus Distribution Regression

1. **Standard Regression:** predict real-valued response Y from a vector-valued covariate $X \in \mathbb{R}^r$
2. **Kernel Regression:** predict real-valued response Y from a covariate $X \in \mathcal{X}$ on which a positive definite (p.d.) kernel exists

Kernel Regression versus Distribution Regression

1. **Standard Regression:** predict real-valued response Y from a vector-valued covariate $X \in \mathbb{R}^r$
2. **Kernel Regression:** predict real-valued response Y from a covariate $X \in \mathcal{X}$ on which a positive definite (p.d.) kernel exists

Distribution regression = kernel regression with $\mathcal{X} = \mathcal{P}(\mathbb{R}^r)$?

Kernel Regression versus Distribution Regression

1. **Standard Regression:** predict real-valued response Y from a vector-valued covariate $X \in \mathbb{R}^r$
2. **Kernel Regression:** predict real-valued response Y from a covariate $X \in \mathcal{X}$ on which a positive definite (p.d.) kernel exists

Distribution regression = kernel regression with $\mathcal{X} = \mathcal{P}(\mathbb{R}^r)$?

- Yes. Finding a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$ is an essential requirement

Kernel Regression versus Distribution Regression

1. **Standard Regression:** predict real-valued response Y from a vector-valued covariate $X \in \mathbb{R}^r$
2. **Kernel Regression:** predict real-valued response Y from a covariate $X \in \mathcal{X}$ on which a positive definite (p.d.) kernel exists

Distribution regression = kernel regression with $\mathcal{X} = \mathcal{P}(\mathbb{R}^r)$?

- Yes. Finding a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$ is an essential requirement
- No. We do not have access to the true samples $\mathbb{P} \in \mathcal{P}(\mathbb{R}^r)$, only

$$\hat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \approx \mathbb{P}_i, \qquad X_i \sim \mathbb{P}$$

Problem set-up

- Distribution ρ on $\mathcal{P}(\mathbb{R}^r) \times Y$, $\rho(\mathbb{P}, y) = \rho(y \mid \mathbb{P})\rho(\mathbb{P})$

Problem set-up

- Distribution ρ on $\mathcal{P}(\mathbb{R}^r) \times Y$, $\rho(\mathbb{P}, y) = \rho(y \mid \mathbb{P})\rho(\mathbb{P})$
- **Expected risk:** $f : \mathcal{P}(\mathbb{R}^r) \to Y$ (meas.)

$$\mathcal{E}(f) = \int_{\mathcal{P}(\mathbb{R}^r) \times Y} (f(\mathbb{P}) - y)^2 d\rho(\mathbb{P}, y)$$

Problem set-up

- Distribution ρ on $\mathcal{P}(\mathbb{R}^r) \times Y$, $\rho(\mathbb{P}, y) = \rho(y \mid \mathbb{P})\rho(\mathbb{P})$
- **Expected risk:** $f : \mathcal{P}(\mathbb{R}^r) \to Y$ (meas.)

$$\mathcal{E}(f) = \int_{\mathcal{P}(\mathbb{R}^r)\times Y} (f(\mathbb{P}) - y)^2 d\rho(\mathbb{P}, y)$$

- **Bayes estimator:** $f_\rho(\mathbb{P}) := \arg\min_f \mathcal{E}(f) = \mathbb{E}[Y \mid \mathbb{P}]$ (unknown)

Problem set-up

- Distribution ρ on $\mathcal{P}(\mathbb{R}^r) \times Y$, $\rho(\mathbb{P}, y) = \rho(y \mid \mathbb{P})\rho(\mathbb{P})$
- **Expected risk:** $f : \mathcal{P}(\mathbb{R}^r) \to Y$ (meas.)

$$\mathcal{E}(f) = \int_{\mathcal{P}(\mathbb{R}^r)\times Y} (f(\mathbb{P}) - y)^2 d\rho(\mathbb{P}, y)$$

- **Bayes estimator:** $f_\rho(\mathbb{P}) := \arg\min_f \mathcal{E}(f) = \mathbb{E}[Y \mid \mathbb{P}]$ (unknown)
- **First stage sampling:** $(\mathbb{P}_t, y_t)_{t=1}^T \sim^{i.i.d} \rho$

Problem set-up

- Distribution ρ on $\mathcal{P}(\mathbb{R}^r) \times Y$, $\rho(\mathbb{P}, y) = \rho(y \mid \mathbb{P})\rho(\mathbb{P})$
- **Expected risk:** $f : \mathcal{P}(\mathbb{R}^r) \to Y$ (meas.)

$$\mathcal{E}(f) = \int_{\mathcal{P}(\mathbb{R}^r)\times Y} (f(\mathbb{P}) - y)^2 d\rho(\mathbb{P}, y)$$

- **Bayes estimator:** $f_\rho(\mathbb{P}) := \arg\min_f \mathcal{E}(f) = \mathbb{E}[Y \mid \mathbb{P}]$ (unknown)
- **First stage sampling:** $(\mathbb{P}_t, y_t)_{t=1}^T \sim^{i.i.d} \rho$
- **Second stage sampling:** $x_{t,i} \sim^{i.i.d.} \mathbb{P}_t$ $(1 \leq t \leq T, 1 \leq i \leq n)$
  Dataset $\mathcal{D} = \{((x_{t,i})_{i=1}^n, y_t)\}_{t=1}^T$

Problem set-up

- Distribution ρ on $\mathcal{P}(\mathbb{R}^r) \times Y$, $\rho(\mathbb{P}, y) = \rho(y \mid \mathbb{P})\rho(\mathbb{P})$
- **Expected risk:** $f : \mathcal{P}(\mathbb{R}^r) \to Y$ (meas.)

$$\mathcal{E}(f) = \int_{\mathcal{P}(\mathbb{R}^r)\times Y} (f(\mathbb{P}) - y)^2 d\rho(\mathbb{P}, y)$$

- **Bayes estimator:** $f_\rho(\mathbb{P}) := \arg\min_f \mathcal{E}(f) = \mathbb{E}[Y \mid \mathbb{P}]$ (unknown)
- **First stage sampling:** $(\mathbb{P}_t, y_t)_{t=1}^T \sim^{i.i.d} \rho$
- **Second stage sampling:** $x_{t,i} \sim^{i.i.d.} \mathbb{P}_t$ $(1 \leq t \leq T, 1 \leq i \leq n)$
  Dataset $\mathcal{D} = \{((x_{t,i})_{i=1}^n, y_t)\}_{t=1}^T$
- **Estimator:** $\hat{f}_\mathcal{D} : \mathcal{P}(\mathbb{R}^r) \longrightarrow Y$
- **Generalisation error:** $\mathcal{E}(\hat{f}_\mathcal{D}) - \mathcal{E}(\hat{f}_\rho)$ small

Kernel Distribution Regression

KDR - Kernel Distribution Regression

Consider p.d. kernel $K : \mathcal{P}(\mathbb{R}^r) \times \mathcal{P}(\mathbb{R}^r) \to \mathbb{R}_+$ with RKHS $\mathcal{H}_K$

$$\mathcal{E}_{T,n,\lambda}(f) := \frac{1}{T}\sum_{t=1}^{T}\left(f\left(\hat{\mathbb{P}}_{t,n}\right) - y_t\right)^2 + \lambda\|f\|^2_{\mathcal{H}_K}$$

$$f_{\mathcal{D},\lambda} := \arg\min_{f\in\mathcal{H}_K} \mathcal{E}_{T,n,\lambda}(f) = (y_1,\dots,y_T)(K_T + \lambda T I_T)^{-1} k_{\mathbb{P}}$$

$[K_T]_{t,l} = K(\hat{\mathbb{P}}_{t,n}, \hat{\mathbb{P}}_{l,n}) \in \mathbb{R}^{T\times T}$,
$k_{\mathbb{P}} = (K(\mathbb{P}, \hat{\mathbb{P}}_{t,n}),\dots,K(\mathbb{P}, \hat{\mathbb{P}}_{T,n}))^\top \in \mathbb{R}^T$

Kernel Distribution Regression

KDR - Kernel Distribution Regression

Consider p.d. kernel $K : \mathcal{P}(\mathbb{R}^r) \times \mathcal{P}(\mathbb{R}^r) \to \mathbb{R}_+$ with RKHS $\mathcal{H}_K$

$$\mathcal{E}_{T,n,\lambda}(f) := \frac{1}{T} \sum_{t=1}^{T} \left(f\left(\hat{\mathbb{P}}_{t,n}\right) - y_t \right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2$$

$$f_{\mathcal{D},\lambda} := \arg\min_{f \in \mathcal{H}_K} \mathcal{E}_{T,n,\lambda}(f) = (y_1, \ldots, y_T)(K_T + \lambda T I_T)^{-1} k_{\mathbb{P}}$$

$[K_T]_{t,l} = K(\hat{\mathbb{P}}_{t,n}, \hat{\mathbb{P}}_{l,n}) \in \mathbb{R}^{T \times T}$,
$k_{\mathbb{P}} = (K(\mathbb{P}, \hat{\mathbb{P}}_{t,n}), \ldots, K(\mathbb{P}, \hat{\mathbb{P}}_{T,n}))^\top \in \mathbb{R}^T$

How to find a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$?

Distributional kernel

How to find a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$? Intuition: Gaussian kernel.

$$K_{\mathrm{Gauss}}(x, x') = e^{-\gamma \|x - x'\|^2_{\mathbb{R}^r}} \qquad (x, x' \in \mathbb{R}^r)$$

$\|x - x'\|_{\mathbb{R}^r}$ Euclidean distance.

Distributional kernel

How to find a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$? Intuition: Gaussian kernel.

$$K_{\text{Gauss}}(\mathbb{P}, \mathbb{P}') = e^{-\gamma d(\mathbb{P}, \mathbb{P}')^2} \qquad (\mathbb{P}, \mathbb{P}' \in \mathcal{P}(\mathbb{R}^r))$$

$d(\mathbb{P}, \mathbb{P}')$ distance on $\mathcal{P}(\mathbb{R}^r)$.

Distributional kernel

How to find a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$? Intuition: Gaussian kernel.

$$K_{\mathrm{Gauss}}(\mathbb{P}, \mathbb{P}') = e^{-\gamma d(\mathbb{P}, \mathbb{P}')^2} \qquad (\mathbb{P}, \mathbb{P}' \in \mathcal{P}(\mathbb{R}^r))$$

$d(\mathbb{P}, \mathbb{P}')$ distance on $\mathcal{P}(\mathbb{R}^r)$.

Hilbertian distance

K_{Gauss} defines a p.d. kernel if and only if there is a Hilbert space $\mathcal{F}$ and a feature map $\Phi : \mathcal{P}(\mathbb{R}^r) \to \mathcal{F}$ such that

$$d(\mathbb{P}, \mathbb{P}') = \|\Phi(\mathbb{P}) - \Phi(\mathbb{P}')\|_{\mathcal{F}}$$

Distributional kernel

How to find a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$? Intuition: Gaussian kernel.

$$K_{\mathrm{Gauss}}(\mathbb{P}, \mathbb{P}') = e^{-\gamma d(\mathbb{P}, \mathbb{P}')^2} \qquad (\mathbb{P}, \mathbb{P}' \in \mathcal{P}(\mathbb{R}^r))$$

$d(\mathbb{P}, \mathbb{P}')$ distance on $\mathcal{P}(\mathbb{R}^r)$.

Hilbertian distance

K_{Gauss} defines a p.d. kernel if and only if there is a Hilbert space $\mathcal{F}$ and a feature map $\Phi : \mathcal{P}(\mathbb{R}^r) \to \mathcal{F}$ such that

$$d(\mathbb{P}, \mathbb{P}') = \|\Phi(\mathbb{P}) - \Phi(\mathbb{P}')\|_{\mathcal{F}}$$

Examples: Maximum Mean Discrepancy, Hellinger, square root Total variation etc

Distributional kernel

How to find a p.d. kernel on $\mathcal{P}(\mathbb{R}^r)$? Intuition: Gaussian kernel.

$$K_{\mathrm{Gauss}}(\mathbb{P}, \mathbb{P}') = e^{-\gamma d(\mathbb{P}, \mathbb{P}')^2} \qquad (\mathbb{P}, \mathbb{P}' \in \mathcal{P}(\mathbb{R}^r))$$

$d(\mathbb{P}, \mathbb{P}')$ distance on $\mathcal{P}(\mathbb{R}^r)$.

Hilbertian distance

K_{Gauss} defines a p.d. kernel if and only if there is a Hilbert space $\mathcal{F}$ and a feature map $\Phi : \mathcal{P}(\mathbb{R}^r) \to \mathcal{F}$ such that

$$d(\mathbb{P}, \mathbb{P}') = \|\Phi(\mathbb{P}) - \Phi(\mathbb{P}')\|_{\mathcal{F}}$$

Examples: Maximum Mean Discrepancy, Hellinger, square root Total variation etc

What about Optimal Transport distances? The Wasserstein distance is not Hilbertian.

Sliced Wasserstein Kernel

1D Optimal Transport

On $\mathbb{R}$, the Wasserstein distance admits a closed-form:

$$d_{W_2}(\mathbb{P}, \mathbb{P}') = \left(\int_{(0,1)} \left(F_{\mathbb{P}}^{[-1]}(t) - F_{\mathbb{P}'}^{[-1]}(t) \right)^2 dt \right)^{\frac{1}{2}}$$

Sliced Wasserstein Kernel

1D Optimal Transport

On $\mathbb{R}$, the Wasserstein distance admits a closed-form:

$$d_{W_2}(\mathbb{P},\mathbb{P}') = \left(\int_{(0,1)} \left(F_{\mathbb{P}}^{[-1]}(t) - F_{\mathbb{P}'}^{[-1]}(t)\right)^2 dt\right)^{\frac{1}{2}}$$

Sliced Wasserstein distance

On $\mathbb{R}^r$ ($r > 1$), the Sliced-Wasserstein distance is:

$$d_{SW_2}(\mathbb{P},\mathbb{P}') = \left(\int_{\mathbb{S}^{d-1}} d_{W_2}(\theta_{\#}\mathbb{P}, \theta_{\#}\mathbb{P}')^2 d\theta\right)^{\frac{1}{2}}$$

Theoretical Results

Under suitable assumptions, with $\lambda = \max(\frac{1}{\sqrt{T}}, \frac{1}{n^{1/4}})$ we have

$$\mathcal{E}(\hat{f}_{\mathcal{D},\lambda}) - \mathcal{E}(f_\rho) \leq C \left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt[4]{n}} \right) \left(\|f_\rho\|_{\mathcal{H}_K} + 1 \right)$$

More in the paper

- Bounds for general Hilbertian distances
- Universality
- Experiments: strong empirical performances of the Sliced Wasserstein kernel in comparison to MMD-based kernels