Kernel Machines

Alain Celisse

SAMM

Paris 1-Panthéon Sorbonne University

`alain.celisse@univ-paris1.fr`

Lecture 5: Maximum Mean Discrepancy

—

Master 2 Data Science – Centrale Lille, Lille University
Fall 2022

Outline of the lectures

Successive topics of the coming lectures:

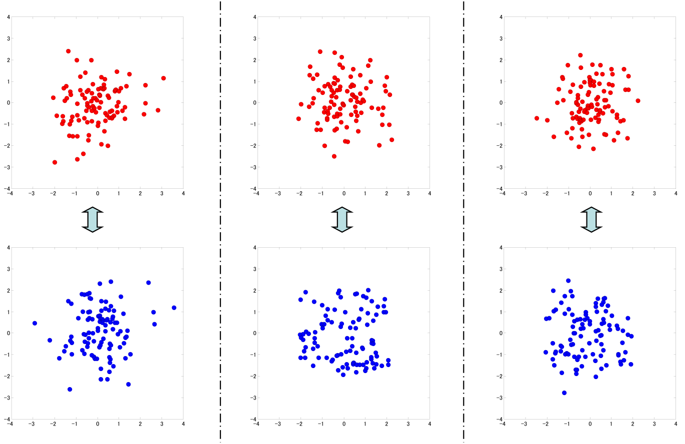
1. Introduction to Kernel methods
2. Support vector classifiers and Kernel methods
3. Extending classical strategies to high dimension
   - KRR/LS-SVMs
   - KPCA
4. Duality gap and KKT conditions
5. Designing reproducing kernels
6. Maximum Mean Discrepancy (MMD) (Today!)

Outline of the lecture

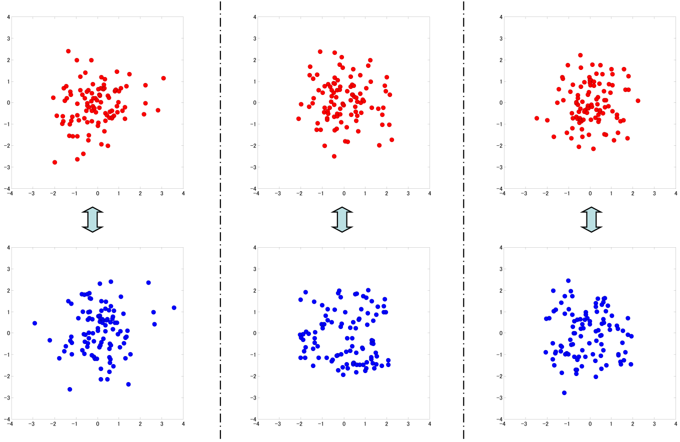
- Distance between Probability distributions
- Maximum Mean Discrepancy (MMD)
- Mean embedding
- Estimation
- Two-sample test

Distance between Probability distributions

Example: Two-sample test



Example: Two-sample test



Question:

Are red and blue samples drawn from the same distribution?

Example: Two-sample test

Two-sample test

$X_1, \ldots, X_n \sim P_X$ and $Y_1, \ldots, Y_m \sim P_Y$, with $P_X = P_Y$?

Example: Two-sample test

Two-sample test

$$X_1, \ldots, X_n \sim P_X \text{ and } Y_1, \ldots, Y_m \sim P_Y, \text{ with } P_X = P_Y?$$

Connections:

- Independence testing
  (comparison between $P_X \otimes P_X$ and $P_{X,Y}$)

Example: Two-sample test

Two-sample test

$$X_1, \ldots, X_n \sim P_X \text{ and } Y_1, \ldots, Y_m \sim P_Y, \text{ with } P_X = P_Y?$$

Connections:

- Independence testing
  (comparison between $P_X \otimes P_X$ and $P_{X,Y}$)
- Novelty/Anomaly detection (on-line)

Example: Two-sample test

Two-sample test

$$X_1, \dots, X_n \sim P_X \text{ and } Y_1, \dots, Y_m \sim P_Y, \text{ with } P_X = P_Y?$$

Connections:

- Independence testing
  (comparison between $P_X \otimes P_X$ and $P_{X,Y}$)
- Novelty/Anomaly detection (on-line)
- Change-point detection (see the next slides!)

Example: Two-sample test

Two-sample test

$$X_1, \ldots, X_n \sim P_X \text{ and } Y_1, \ldots, Y_m \sim P_Y, \text{ with } P_X = P_Y?$$

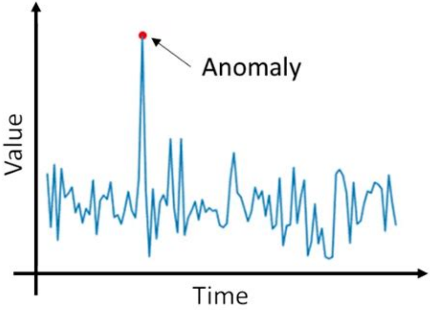
Connections:

- Independence testing
  (comparison between $P_X \otimes P_X$ and $P_{X,Y}$)
- Novelty/Anomaly detection (on-line)
- Change-point detection (see the next slides!)

Key ingredient

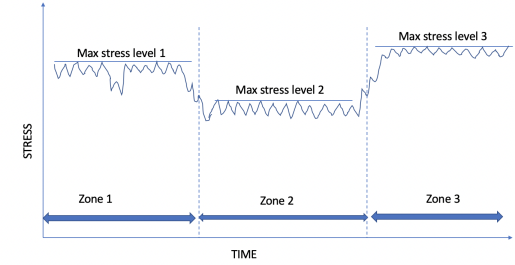
Requires a "distance" between P_X and P_Y

Anomaly detection: Online scenario

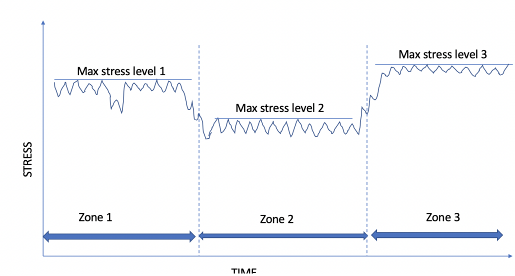


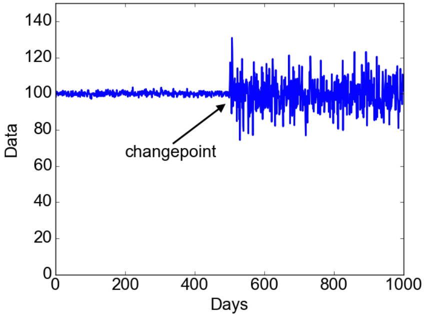
- Reference distribution until time $t_0 > 0$
- **Goal:** From $t > t_0$, detecting (in real-time) potential shifts in the distribution

Change-point detection: Offline scenario



Change-point detection: Offline scenario





Distance between Probability distributions

Kolmogorov-Smirnov distance (KS-distance)

- $X_1, \ldots, X_n$, and $Y_1, \ldots, Y_m \in \mathbb{R}$
- Cumulative distribution functions are characteristic:

Theorem

$$F(t) = \mathbb{P}[X \leq t] = G(t) = \mathbb{P}[Y \leq t], \quad \forall t \in \mathbb{R}$$

implies that $P_X = P_Y$

KS-distance between P and Q

Definition (KS-distance)

$$\begin{aligned} d_{KS}(P, Q) &= \sup_{t \in \mathbb{R}} |F(t) - G(t)| = \|F - G\|_\infty \\ &= \sup_{t \in \mathbb{R}} \left| \mathbb{E}\left[\mathbb{1}_{(X \leq t)} \right] - \mathbb{E}\left[\mathbb{1}_{(Y \leq t)} \right] \right| \\ &= \sup_{f \in \mathcal{F}} |\mathbb{E}[f(X) - f(Y)]| \end{aligned}$$

$$\mathcal{F} = \left\{ f = \mathbb{1}_{]-\infty, t]} \mid t \in \mathbb{R} \right\}$$

Other famous metrics

Integral probability metrics

For some class $\mathcal{F}$,

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

Other famous metrics

Integral probability metrics

For some class $\mathcal{F}$,

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

Total variation (TV) distance

$$\mathcal{F} = \mathcal{F}_{TV} = \{ f \in \mathcal{M}(\mathcal{X}) \mid \|f\|_\infty \leq 1 \}$$

Other famous metrics

Integral probability metrics

For some class $\mathcal{F}$,

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

Total variation (TV) distance

$$\mathcal{F} = \mathcal{F}_{TV} = \{ f \in \mathcal{M}(\mathcal{X}) \mid \|f\|_\infty \leq 1 \}$$

Wasserstein (W) distance

$$\mathcal{F} = \mathcal{F}_W = \left\{ f \in \mathcal{M}(\mathcal{X}) \mid \|f\|_L = \sup_{x \neq y} \left| \frac{f(x) - f(y)}{\rho(x, y)} \right| \leq 1 \right\}$$

Other famous metrics

Integral probability metrics

For some class $\mathcal{F}$,

$$d_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

Total variation (TV) distance

$$\mathcal{F} = \mathcal{F}_{TV} = \{ f \in \mathcal{M}(\mathcal{X}) \mid \|f\|_\infty \leq 1 \}$$

Wasserstein (W) distance

$$\mathcal{F} = \mathcal{F}_W = \left\{ f \in \mathcal{M}(\mathcal{X}) \mid \|f\|_L = \sup_{x \neq y} \left| \frac{f(x) - f(y)}{\rho(x,y)} \right| \leq 1 \right\}$$

L^q-distance

$$\mathcal{F} = \mathcal{F}_q = \left\{ f \in \mathcal{M}(\mathcal{X}) \mid \|f\|_q^q = \int_{\mathcal{X}} |f|^q \, d\lambda \leq 1 \right\}$$

$$(1 \leq q < +\infty)$$

Main limitations

Computations

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

- Previous metrics defined from a supremum

Main limitations

Computations

$$d_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\, f(X) - f(Y)\,\right]\right|$$

- Previous metrics defined from a supremum
- $\mathcal{F}$ is infinite

Main limitations

Computations

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

- Previous metrics defined from a supremum
- $\mathcal{F}$ is infinite
- $d_{\mathcal{F}}(P, Q)$ usually difficult to compute

Main limitations

Computations

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

- Previous metrics defined from a supremum
- $\mathcal{F}$ is infinite
- $d_{\mathcal{F}}(P, Q)$ usually difficult to compute

Richness of $\mathcal{F}$

- The richness of $\mathcal{F}$ determines the properties of $d_{\mathcal{F}}(\cdot, \cdot)$

Main limitations

Computations

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

- Previous metrics defined from a supremum
- $\mathcal{F}$ is infinite
- $d_{\mathcal{F}}(P, Q)$ usually difficult to compute

Richness of $\mathcal{F}$

- The richness of $\mathcal{F}$ determines the properties of $d_{\mathcal{F}}(\cdot, \cdot)$
- **Ex:**
  Is the next set characteristic?

$$\mathcal{F} = \left\{ \theta \left| \cdot \right|^2 \mid \theta \in \mathbb{R} \right\}$$

Main limitations

Computations

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

- Previous metrics defined from a supremum
- $\mathcal{F}$ is infinite
- $d_{\mathcal{F}}(P, Q)$ usually difficult to compute

Richness of $\mathcal{F}$

- The richness of $\mathcal{F}$ determines the properties of $d_{\mathcal{F}}(\cdot, \cdot)$
- **Ex:**
  Is the next set characteristic?

$$\mathcal{F} = \left\{ \theta \left| \cdot \right|^2 \mid \theta \in \mathbb{R} \right\}$$

$\longrightarrow$ Infinite but too much thin for spanning a metric!

Maximum Mean Discrepancy (MMD)

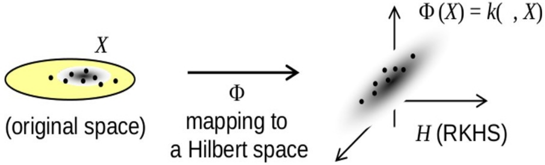
Psd kernel and extended features

Mapping data from $\mathcal{X}$ to $\mathcal{H}$

- $k(\cdot,\cdot)$: psd kernel (repoducing)
- $x \in \mathbb{R}^d \mapsto k_x = k(x,\cdot) \in \mathcal{H}$: canonical feature map from $\mathbb{R}^d$ to $\mathcal{H}$ with

$$k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}}$$

- $k_x = \phi(x) \in \mathcal{H}$: new "observation"



Reproducing kernel and RKHS

Theorem (From psd to reproducing kernel)

Any psd kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives rise to a unique Hilbert space endowed with the scalar product

$$k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}}, \qquad \forall x, y \in \mathcal{X}$$

and such that

Reproducing kernel and RKHS

Theorem (From psd to reproducing kernel)

Any psd kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives rise to a unique Hilbert space endowed with the scalar product

$$k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}}\,, \qquad \forall x, y \in \mathcal{X}$$

and such that

1. *$\mathcal{H}$ contains all functions $k_x : x \mapsto k(x,\cdot)$, for all $x \in \mathcal{X}$*

Reproducing kernel and RKHS

Theorem (From psd to reproducing kernel)

Any psd kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives rise to a unique Hilbert space endowed with the scalar product

$$k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}}\,, \qquad \forall x, y \in \mathcal{X}$$

and such that

1. *$\mathcal{H}$ contains all functions $k_x : x \mapsto k(x,\cdot)$, for all $x \in \mathcal{X}$*
2. *$\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$,*

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} \qquad \textit{(Reproducing property)}$$

Reproducing kernel and RKHS

Theorem (From psd to reproducing kernel)

Any psd kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives rise to a unique Hilbert space endowed with the scalar product

$$k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}}, \qquad \forall x, y \in \mathcal{X}$$

and such that

1. *$\mathcal{H}$ contains all functions $k_x : x \mapsto k(x,\cdot)$, for all $x \in \mathcal{X}$*
2. *$\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$,*

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} \qquad \text{(Reproducing property)}$$

Then,

- *$\mathcal{H} = \mathcal{H}_k$: Reproducing Kernel Hilbert Space (RKHS) associated with k.*

Reproducing kernel and RKHS

Theorem (From psd to reproducing kernel)

Any psd kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives rise to a unique Hilbert space endowed with the scalar product

$$k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}}\,, \qquad \forall x,y \in \mathcal{X}$$

and such that

1. *$\mathcal{H}$ contains all functions $k_x : x \mapsto k(x,\cdot)$, for all $x \in \mathcal{X}$*
2. *$\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$,*

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} \qquad \text{(Reproducing property)}$$

Then,

- *$\mathcal{H} = \mathcal{H}_k$: Reproducing Kernel Hilbert Space (RKHS) associated with k.*
- *k: Reproducing kernel of $\mathcal{H}$.*

Maximum Mean Discrepancy (MMD) (1/2)

Definition (MMD)

(Gretton, Fukumizu, Sriperumbudur,...)

$$MMD_{\mathcal{H}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}[f(X) - f(Y)]|$$

Maximum Mean Discrepancy (MMD) (1/2)

Definition (MMD)

(Gretton, Fukumizu, Sriperumbudur,. . .)

$$MMD_{\mathcal{H}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

where

$$\mathcal{F} = \{ f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1 \}$$

Maximum Mean Discrepancy (MMD) (2/2)

$$MMD_{\mathcal{H}}(P,Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}\left[f(X) - f(Y) \right]|$$

Maximum Mean Discrepancy (MMD) (2/2)

$$MMD_{\mathcal{H}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}[f(X) - f(Y)]|$$

Calculating the MMD

Maximum Mean Discrepancy (MMD) (2/2)

$$MMD_{\mathcal{H}}(P,Q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\, f(X) - f(Y)\,\right]\right|$$

Calculating the MMD

$$\begin{aligned}
MMD_{\mathcal{H}}(P,Q) &= \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X,Y}\left[\, f(X) - f(Y)\,\right]\right| \\
&= \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X,Y}\left[\, \langle f, k_X - k_Y\rangle_{\mathcal{H}}\,\right]\right| \\
&= \sup_{f\in\mathcal{F}} \left|\langle f, \mathbb{E}_{X,Y}\left[\, k_X - k_Y\,\right]\rangle_{\mathcal{H}}\right| \\
\text{``} &= \left\|\mathbb{E}_{X,Y}\left[\, k_X - k_Y\,\right]\right\|_{\mathcal{H}}\text{''}
\end{aligned}$$

Maximum Mean Discrepancy (MMD) (2/2)

$$MMD_{\mathcal{H}}(P,Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}\left[f(X) - f(Y) \right]|$$

Calculating the MMD

$$\begin{aligned} MMD_{\mathcal{H}}(P,Q) &= \sup_{f\in\mathcal{F}} |\mathbb{E}_{X,Y}\left[f(X) - f(Y) \right]| \\ &= \sup_{f\in\mathcal{F}} |\mathbb{E}_{X,Y}\left[\langle f, k_X - k_Y \rangle_{\mathcal{H}} \right]| \\ &= \sup_{f\in\mathcal{F}} \left| \langle f, \mathbb{E}_{X,Y}\left[k_X - k_Y \right] \rangle_{\mathcal{H}} \right| \\ \text{``} &= \|\mathbb{E}_{X,Y}\left[k_X - k_Y \right]\|_{\mathcal{H}}\text{''} \end{aligned}$$

Remarks:

- No supremum anymore!

Maximum Mean Discrepancy (MMD) (2/2)

$$MMD_{\mathcal{H}}(P,Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}\,[\, f(X) - f(Y)\,]|$$

Calculating the MMD

$$\begin{aligned}
MMD_{\mathcal{H}}(P,Q) &= \sup_{f \in \mathcal{F}} |\mathbb{E}_{X,Y}\,[\, f(X) - f(Y)\,]| \\
&= \sup_{f \in \mathcal{F}} |\mathbb{E}_{X,Y}\,[\, \langle f, k_X - k_Y \rangle_{\mathcal{H}}\,]| \\
&= \sup_{f \in \mathcal{F}} \left| \langle f, \mathbb{E}_{X,Y}\,[\, k_X - k_Y\,] \rangle_{\mathcal{H}} \right| \\
\text{``} &= \| \mathbb{E}_{X,Y}\,[\, k_X - k_Y\,] \|_{\mathcal{H}}\text{''}
\end{aligned}$$

Remarks:

- No supremum anymore!
- Meaning of $\mathbb{E}_{X,Y}\,[\, k_X - k_Y\,]$?

Maximum Mean Discrepancy (MMD) (2/2)

$$MMD_{\mathcal{H}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right|$$

Calculating the MMD

$$\begin{aligned}
MMD_{\mathcal{H}}(P, Q) &= \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X,Y}\left[f(X) - f(Y) \right] \right| \\
&= \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X,Y}\left[\langle f, k_X - k_Y \rangle_{\mathcal{H}} \right] \right| \\
&= \sup_{f \in \mathcal{F}} \left| \langle f, \mathbb{E}_{X,Y}\left[k_X - k_Y \right] \rangle_{\mathcal{H}} \right| \\
\text{``} &= \left\| \mathbb{E}_{X,Y}\left[k_X - k_Y \right] \right\|_{\mathcal{H}}\text{''}
\end{aligned}$$

Remarks:

- No supremum anymore!
- Meaning of $\mathbb{E}_{X,Y}\left[k_X - k_Y \right]$? $\rightarrow$ Mean element

Mean Embedding/Mean Element

Definition (Mean element)

- k: psd kernel on $\mathcal{X}$
- Assume: $\mathbb{E}\left[\sqrt{k(X,X)}\right] = \mathbb{E}\left[\, \|k_X\|_{\mathcal{H}} \,\right] < +\infty$

Mean Embedding/Mean Element

Definition (Mean element)

- k: psd kernel on $\mathcal{X}$
- Assume: $\mathbb{E}\left[\sqrt{k(X,X)}\right] = \mathbb{E}\left[\,\|k_X\|_{\mathcal{H}}\,\right] < +\infty$

Then, there exists a unique element $\mu_P \in \mathcal{H}$ such that

$$\langle \mu_P, f \rangle_{\mathcal{H}} = \mathbb{E}[\,\langle k_X, f\rangle_{\mathcal{H}}\,], \qquad \forall f \in \mathcal{H}$$

Mean Embedding/Mean Element

Definition (Mean element)

- k: psd kernel on $\mathcal{X}$
- Assume: $\mathbb{E}\left[\sqrt{k(X,X)}\right] = \mathbb{E}\left[\,\|k_X\|_{\mathcal{H}}\,\right] < +\infty$

Then, there exists a unique element $\mu_P \in \mathcal{H}$ such that

$$\langle \mu_P, f\rangle_{\mathcal{H}} = \mathbb{E}[\,\langle k_X, f\rangle_{\mathcal{H}}\,], \qquad \forall f \in \mathcal{H}$$

Remark: $\mathbb{E}\left[\sqrt{k(X,X)}\right] < +\infty$ true for any bounded kernel (e.g. normalized kernel)

Mean Embedding/Mean Element

Definition (Mean element)

- k: psd kernel on $\mathcal{X}$
- Assume: $\mathbb{E}\left[\sqrt{k(X,X)}\right] = \mathbb{E}\left[\,\|k_X\|_{\mathcal{H}}\,\right] < +\infty$

Then, there exists a unique element $\mu_P \in \mathcal{H}$ such that

$$\langle \mu_P, f\rangle_{\mathcal{H}} = \mathbb{E}[\,\langle k_X, f\rangle_{\mathcal{H}}\,], \qquad \forall f \in \mathcal{H}$$

Remark: $\mathbb{E}\left[\sqrt{k(X,X)}\right] < +\infty$ true for any bounded kernel (e.g. normalized kernel)

Proof.

Mean Embedding/Mean Element

Definition (Mean element)

- k: psd kernel on $\mathcal{X}$
- Assume: $\mathbb{E}\left[\sqrt{k(X,X)}\right] = \mathbb{E}\left[\,\|k_X\|_{\mathcal{H}}\,\right] < +\infty$

Then, there exists a unique element $\mu_P \in \mathcal{H}$ such that

$$\langle \mu_P, f\rangle_{\mathcal{H}} = \mathbb{E}[\,\langle k_X, f\rangle_{\mathcal{H}}\,], \qquad \forall f \in \mathcal{H}$$

Remark: $\mathbb{E}\left[\sqrt{k(X,X)}\right] < +\infty$ true for any bounded kernel (e.g. normalized kernel)

Proof.

$f \mapsto \mathbb{E}\left[\,\langle k_X, f\rangle_{\mathcal{H}}\,\right]$ is a bounded linear form over $\mathcal{H}$ since

$$\left|\mathbb{E}\left[\,\langle k_X, f\rangle_{\mathcal{H}}\,\right]\right| \leq \mathbb{E}\left[\,\|k_X\|_{\mathcal{H}}\,\right] \|f\|_{\mathcal{H}}$$

Riesz's theorem yields the existence and unicity $\square$

Basic properties of the mean element

Theorem

- $\mu_P = \mathbb{E}_X [\, k_X \,] \in \mathcal{H}$

Basic properties of the mean element

Theorem

- $\mu_P = \mathbb{E}_X\left[\, k_X \,\right] \in \mathcal{H}$

- *For all* $x \in \mathcal{X}$,

$$\mu_P(x) = \mathbb{E}_X\left[\, k_X(x) \,\right] = \mathbb{E}_X\left[\, k(X, x) \,\right] = \int_{\mathcal{X}} k(u, x) dP_X(u)$$

Basic properties of the mean element

Theorem

- $\mu_P = \mathbb{E}_X\left[\, k_X \,\right] \in \mathcal{H}$

- *For all* $x \in \mathcal{X}$,

$$\mu_P(x) = \mathbb{E}_X\left[\, k_X(x) \,\right] = \mathbb{E}_X\left[\, k(X, x) \,\right] = \int_{\mathcal{X}} k(u, x) dP_X(u)$$

- $\mathbb{E}_X\left[\, k_X \,\right] + \mathbb{E}_X\left[\, k_Y \,\right] = \mu_P + \mu_Q$

Back to the. . . MMD

MMD and mean element

$$\begin{aligned} MMD_{\mathcal{H}}(P,Q) &= \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\, f(X) - f(Y)\,\right]\right| \\ &= \sup_{f\in\mathcal{F}} \left|\langle f, \mathbb{E}\left[\, k_X - k_Y \,\right]\rangle_{\mathcal{H}}\right| \\ &= \left\|\mathbb{E}\left[\, k_X - k_Y \,\right]\right\|_{\mathcal{H}} \\ &= \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}} \end{aligned}$$

Back to the. . . MMD

MMD and mean element

$$
\begin{aligned}
MMD_{\mathcal{H}}(P, Q) &= \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right| \\
&= \sup_{f \in \mathcal{F}} \left| \langle f, \mathbb{E}\left[k_X - k_Y \right] \rangle_{\mathcal{H}} \right| \\
&= \left\| \mathbb{E}\left[k_X - k_Y \right] \right\|_{\mathcal{H}} \\
&= \left\| \mu_P - \mu_Q \right\|_{\mathcal{H}}
\end{aligned}
$$

The "distance" between P and Q translates into a difference between the mean elements of μ_P and μ_Q in $\|\cdot\|_{\mathcal{H}}$

Back to the...MMD

MMD and mean element

$$
\begin{aligned}
MMD_{\mathcal{H}}(P, Q) &= \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(X) - f(Y) \right] \right| \\
&= \sup_{f \in \mathcal{F}} \left| \langle f, \mathbb{E}\left[k_X - k_Y \right] \rangle_{\mathcal{H}} \right| \\
&= \left\| \mathbb{E}\left[k_X - k_Y \right] \right\|_{\mathcal{H}} \\
&= \left\| \mu_P - \mu_Q \right\|_{\mathcal{H}}
\end{aligned}
$$

The "distance" between P and Q translates into a difference between the mean elements of μ_P and μ_Q in $\|\cdot\|_{\mathcal{H}}$

Remark:

$$
P = Q \quad \Rightarrow \quad \|\mu_P - \mu_Q\|_{\mathcal{H}} = 0 \quad \Leftrightarrow \quad \mu_P = \mu_Q
$$

Back to the... MMD

MMD and mean element

$$
\begin{aligned}
MMD_{\mathcal{H}}(P,Q) &= \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\, f(X) - f(Y)\,\right]\right| \\
&= \sup_{f\in\mathcal{F}} \left|\langle f, \mathbb{E}\left[\, k_X - k_Y \,\right]\rangle_{\mathcal{H}}\right| \\
&= \left\|\mathbb{E}\left[\, k_X - k_Y \,\right]\right\|_{\mathcal{H}} \\
&= \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}}
\end{aligned}
$$

The "distance" between P and Q translates into a difference between the mean elements of μ_P and μ_Q in $\|\cdot\|_{\mathcal{H}}$

Remark:

$$
P = Q \quad \Rightarrow \quad \|\mu_P - \mu_Q\|_{\mathcal{H}} = 0 \quad \Leftrightarrow \quad \mu_P = \mu_Q
$$

Warning: The converse is not true in general!

Mean element/Mean embedding

Polynomial kernel and mean element

Mean element

For $X \sim P$

$$\mu_P = \mathbb{E}\left[\, k_X \,\right] = \mathbb{E}\left[\, k(X, \cdot) \,\right] \in \mathcal{H}$$

$$\Rightarrow \quad \mu_P(t) = \mathbb{E}\left[\, k(X, t) \,\right], \quad \forall t \in \mathcal{X}$$

Polynomial kernel and mean element

Mean element

For $X \sim P$

$$\begin{aligned} & \mu_P = \mathbb{E}[\, k_X \,] = \mathbb{E}[\, k(X, \cdot) \,] \in \mathcal{H} \\ \Rightarrow \quad & \mu_P(t) = \mathbb{E}[\, k(X, t) \,], \quad \forall t \in \mathcal{X} \end{aligned}$$

Examples: Polynomial kernels

- $k(X, t) = (Xt)^d$, for $X, t \in \mathbb{R}$ and $d \in \mathbb{N}^*$

$$\mu_P(t) = \mathbb{E}\left[(Xt)^d \right] = \mathbb{E}\left[X^d \right] t^d$$

$\longrightarrow$ only involves the dth moment.

Polynomial kernel and mean element

Mean element

For $X \sim P$

$$\mu_P = \mathbb{E}\,[\,k_X\,] = \mathbb{E}\,[\,k(X,\cdot)\,] \in \mathcal{H}$$
$$\Rightarrow \quad \mu_P(t) = \mathbb{E}\,[\,k(X,t)\,]\,, \quad \forall t \in \mathcal{X}$$

Examples: Polynomial kernels

- $k(X,t) = (Xt)^d$, for $X, t \in \mathbb{R}$ and $d \in \mathbb{N}^*$

$$\mu_P(t) = \mathbb{E}\left[(Xt)^d\right] = \mathbb{E}\left[X^d\right] t^d$$

$\longrightarrow$ only involves the dth moment.

- $k(X,t) = (Xt+c)^d$, for $X, t \in \mathbb{R}$ and $d \in \mathbb{N}^*$ $c > 0$

$$\mu_P(t) = \mathbb{E}\left[(Xt+c)^d\right] = \sum_{i=0}^{d} \binom{d}{i} \mathbb{E}\left[X^i\right] t^i c^{d-i}$$

$\longrightarrow$ involves all moments up to d!

Taylor expansion and mean element

For $X \sim P$

$$\mu_P(t) = \mathbb{E}\left[k(X,t) \right], \quad \forall t \in \mathcal{X}$$

Examples: Taylor expansion

Assume

$$k(X,t) = \sum_{i=0}^{+\infty} a_i X^i t^i, \qquad \forall X, t \in \mathcal{X}$$

Taylor expansion and mean element

For $X \sim P$

$$\mu_P(t) = \mathbb{E}\left[\, k(X,t) \,\right], \quad \forall t \in \mathcal{X}$$

Examples: Taylor expansion

Assume

$$k(X,t) = \sum_{i=0}^{+\infty} a_i X^i t^i, \qquad \forall X, t \in \mathcal{X}$$

Then

$$\mu_P(t) = \sum_{i=0}^{+\infty} a_i \mathbb{E}\left[\, X^i \,\right] t^i,$$

Taylor expansion and mean element

For $X \sim P$

$$\mu_P(t) = \mathbb{E}\left[\, k(X,t) \,\right], \quad \forall t \in \mathcal{X}$$

Examples: Taylor expansion

Assume

$$k(X,t) = \sum_{i=0}^{+\infty} a_i X^i t^i, \qquad \forall X, t \in \mathcal{X}$$

Then

$$\mu_P(t) = \sum_{i=0}^{+\infty} a_i \mathbb{E}\left[\, X^i \,\right] t^i,$$

Remark:

- The exponential kernel
- All moments involved within the mean element
- Construct kernel with prescribed moments...

Characteristic functions and kernels

Laplace transform

$$t \in \mathcal{X} \mapsto \mathcal{L}_P(t) = \mathbb{E}\left[e^{-\langle X, t \rangle} \right]$$

- $\mathcal{L}_P$ is characteristic of the probability distribution P

$$\mathcal{L}_P = \mathcal{L}_Q \quad \Leftrightarrow \qquad P = Q$$

- Since $t \mapsto \mathcal{L}_P(t) = \mu_P(t)$ with the exponential kernel, we have

$$\mu_P = \mu_Q \quad \Rightarrow \quad P = Q$$

Remark:
Analogy with the characteristic function

$$t \in \mathcal{X} \mapsto \phi_P(t) = \mathbb{E}\left[e^{i\langle X, t \rangle} \right]$$

Characteristic kernel

Definition (Characteristic kernel)

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is characteristic if the mapping $P \mapsto \mu_P$ is injective that is,

$$\mu_P = \mu_Q \quad \Rightarrow \quad P = Q$$

Remark:

- $P \mapsto \mu_P$ justifies the name "mean embedding"
- characteristicity through mean element
- the converse inequality holds always true!
  $\to P \mapsto \mu_P$: one-to-one mapping (characteristic kernel)

Characteristic kernel properties

Proposition

The following claims are equivalent:

1. k*: characteristic*
2. $P \mapsto \mu_P$*: injective*
3. $(\mathbb{E}[f(X)] = \mathbb{E}[f(Y)], \quad \forall f \in \mathcal{H}) \Rightarrow P = Q$

Proof.

Do it! $\square$

Describing Characteristic kernels

General domain $\mathcal{X}$

Theorem (Functional description)

With a bounded psd kernel k, the next two claims are equivalent

1. *k is characteristic*
2. *$\mathcal{H} + \mathbb{R}$ dense in all $L^2(R)$, for any probability measure R*

Ex:

Gaussian and Laplace kernels

$\longrightarrow$ Difficult to check in general!

Describing Characteristic kernels (Cont'd)

Theorem (Translation-invariant kernels)

- *k: bounded psd kernel on $\mathbb{R}^d$ with $k(x, y) = h(x - y)$*
- *$h(\cdot)$: bounded, continuous, positive and definite on $\mathbb{R}^d$*

Then Bochner's theorem yields that h is the Fourier transform of a finite non-negative Borel measure Λ that is,

$$h(t) = \int_{\mathbb{R}^d} e^{-i\langle t, w\rangle} d\Lambda(w).$$

Moreover, the next two claims are equivalent

1. *k is characteristic*
2. *$Supp(\Lambda) = \mathbb{R}^d$*

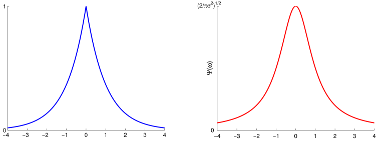
Ex:

Gaussian and Laplace, Matern class, B_{2n+1}-splines,... kernels

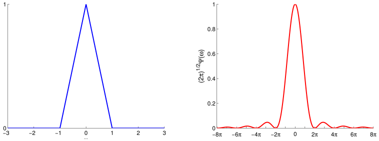
Characteristic kernel: Examples

Characteristic kernels

- Gaussian: $h(t) = e^{-t^2/(2\sigma^2)}$, $\mathcal{F}(h)(u) = \sigma e^{-\sigma^2 u^2/2}$
- Laplace: $h(t) = e^{-|t|\sigma}$, $\mathcal{F}(h)(u) = \sqrt{2/\pi}\frac{\sigma}{\sigma^2+u^2}$



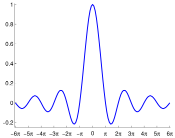
- B_1-Spline kernel: $h(t) = (1-|t|)\mathbb{1}_{[-1,1]}(t)$, $\mathcal{F}(h)(u) = \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{\sin(u/2)^2}{u^2}$

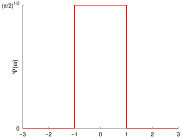


Non characteristic kernels: Examples

Non characteristic kernels

- Sinc kernel: $h(t) = \frac{\sin(\sigma t)}{t}$, $\mathcal{F}(h)(u) = \sqrt{\frac{\pi}{2}}\mathbb{1}_{[-\sigma,\sigma]}(u)$





- Periodic functions h on $\mathbb{R}^d$ in full generality

Back to MMD

Characteristic kernel

With a characteristic kernel,

$$P = Q \quad \Leftrightarrow \quad \mu_P = \mu_Q$$

MMD as a distance over probability measures

With a characteristic kernel,

$$P = Q \quad \Leftrightarrow \quad MMD_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}} = 0$$

Conclusion:

The two claims are equivalent:

- MMD_k: distance over probability distributions
- k: characteristic

Two-sample test

Testing for two fixed populations

Two-sample test

$$X_1, \ldots, X_n \sim P \text{ and } Y_1, \ldots, Y_m \sim Q, \text{ with } P = Q?$$

Statistical test

- Null hypothesis:

$$\mathrm{H}_0 : P = Q \text{ (no change)}$$

- Alternative:

$$\mathrm{H}_1 : P \neq Q$$

- **Oracle** Test Statistic and Rejection region:

$$T = MMD_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}$$
$$\mathcal{R} = \{MMD_k(P, Q) > 0\}$$

  - Should be zero under H_0 and away from 0 otherwise
  - To be estimated...

Designing an estimator

Estimating the mean elements

$$\mu_P = \mathbb{E}\left[\, k(X, \cdot) \,\right] \quad \approx \quad \widehat{\mu}_P = \frac{1}{n} \sum_{i=1}^{n} k(X_i, \cdot)$$

$$\mu_Q = \mathbb{E}\left[\, k(Y, \cdot) \,\right] \quad \approx \quad \widehat{\mu}_Q = \frac{1}{m} \sum_{j=1}^{m} k(Y_j, \cdot)$$

First plug-in estimator of the MMD

$$\begin{aligned}
&\widehat{MMD}^2_{b,k}(P, Q) \\
&= \|\widehat{\mu}_P - \widehat{\mu}_Q\|^2_{\mathcal{H}} \\
&= \|\widehat{\mu}_P\|^2_{\mathcal{H}} + \|\widehat{\mu}_Q\|^2_{\mathcal{H}} - 2\,\langle \widehat{\mu}_P, \widehat{\mu}_Q \rangle_{\mathcal{H}} \\
&= \frac{\sum_{i,j=1}^{n} k(X_i, X_j)}{n^2} + \frac{\sum_{p,q=1}^{m} k(Y_p, Y_q)}{m^2} - 2\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j)}{mn}
\end{aligned}$$

Bias of the plug-in estimator

Covariance operator

Definition

The covariance operator associated with X is the unique operator Σ^X from $\mathcal{H}$ to $\mathcal{H}$ such that

$$\left\langle \Sigma^X f, g \right\rangle_{\mathcal{H}} = \mathbb{E}\left[f(X) \cdot g(X) \right], \qquad \forall f, g \in \mathcal{H}$$

Remark:

From the reproducing property:

$$\begin{aligned} \mathbb{E}\left[\langle f, k_X \rangle_{\mathcal{H}} \langle g, k_X \rangle_{\mathcal{H}} \right] &= \mathbb{E}\left[\langle (k_X \otimes k_X) f, g \rangle_{\mathcal{H}} \right] \\ &= \langle \mathbb{E}\left[k_X \otimes k_X \right] f, g \rangle_{\mathcal{H}} \end{aligned}$$

Theorem

$$\mathbb{E}\left[\widehat{MMD}^2_{b,k}(P, Q) \right]$$

$$= MMD_k^2(P, Q) + \frac{\mathrm{Tr}\left(\Sigma^X\right) - \|\mu_P\|^2_{\mathcal{H}}}{n} + \frac{\mathrm{Tr}\left(\Sigma^Y\right) - \|\mu_Q\|^2_{\mathcal{H}}}{n}$$

Convergence rate

Theorem

- *$\mathcal{K}$: set of candidate kernels*
- *k: kernel such that $\sup_{k\in\mathcal{K}} \|k_x\|_{\mathcal{H}} \leq B$, for all $x \in \mathcal{X}$*

With proba at least $1-\delta$,

$$\left|\widehat{MMD}_{b,k}(P,Q) - MMD_k(P,Q)\right| \leq C\left(\sqrt{\frac{\mathcal{R}ad_m(\mathcal{K})}{m}} + \sqrt{\frac{\mathcal{R}ad_n(\mathcal{K})}{n}}\right) + C'B\left(1+\sqrt{\log\left(\frac{4}{\delta}\right)}\right)\cdot\sqrt{\frac{m+n}{mn}}$$

where the empirical Rademacher complexity is given by

$$\mathcal{R}ad_n(\mathcal{K}) = \mathbb{E}\left[\sup_{k\in\mathcal{K}}\left|\frac{1}{n}\sum_{1\leq i<j\leq n}\epsilon_i\epsilon_j k(X_i,X_j)\right| \mid X_1,\ldots,X_n\right]$$

Designing an estimator: 2nd try ...

$$MMD_k^2(P,Q) = \|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2 - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= \langle \mu_P, \mu_P\rangle_{\mathcal{H}} + \langle \mu_Q, \mu_Q\rangle_{\mathcal{H}} - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= E_{X,X'}\left[\langle k_X, k_{X'}\rangle_{\mathcal{H}}\right] + \mathbb{E}_{Y,Y'}\left[\langle k_Y, k_{Y'}\rangle_{\mathcal{H}}\right] - 2\mathbb{E}_{X,Y}\left[\langle k_X, k_Y\rangle_{\mathcal{H}}\right]$$

Designing an estimator: 2nd try ...

$$MMD_k^2(P,Q) = \|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2 - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= \langle \mu_P, \mu_P\rangle_{\mathcal{H}} + \langle \mu_Q, \mu_Q\rangle_{\mathcal{H}} - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= E_{X,X'}\left[\langle k_X, k_{X'}\rangle_{\mathcal{H}}\right] + \mathbb{E}_{Y,Y'}\left[\langle k_Y, k_{Y'}\rangle_{\mathcal{H}}\right] - 2\mathbb{E}_{X,Y}\left[\langle k_X, k_Y\rangle_{\mathcal{H}}\right]$$

Second plug-in estimator of the MMD

$$\widehat{MMD}^2_{u,k}(P,Q)$$
$$= \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} k(X_i,X_j) + \frac{1}{m(m-1)}\sum_{p\neq q=1}^{m} k(Y_p,Y_q)$$
$$-2\frac{\sum_{i=1}^n\sum_{j=1}^m k(X_i,Y_j)}{mn}$$

Designing an estimator: 2nd try ...

$$MMD_k^2(P,Q) = \|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2 - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= \langle \mu_P, \mu_P\rangle_{\mathcal{H}} + \langle \mu_Q, \mu_Q\rangle_{\mathcal{H}} - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= E_{X,X'}\left[\langle k_X, k_{X'}\rangle_{\mathcal{H}}\right] + \mathbb{E}_{Y,Y'}\left[\langle k_Y, k_{Y'}\rangle_{\mathcal{H}}\right] - 2\mathbb{E}_{X,Y}\left[\langle k_X, k_Y\rangle_{\mathcal{H}}\right]$$

Second plug-in estimator of the MMD

$$\widehat{MMD}^2_{u,k}(P,Q)$$
$$= \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} k(X_i,X_j) + \frac{1}{m(m-1)}\sum_{p\neq q=1}^{m} k(Y_p,Y_q)$$
$$- 2\frac{\sum_{i=1}^n \sum_{j=1}^m k(X_i,Y_j)}{mn}$$

Theorem

$$\mathbb{E}\left[\widehat{MMD}^2_{u,k}(P,Q)\right] = MMD_k^2(P,Q)$$

Designing an estimator: 2nd try . . .

$$MMD_k^2(P,Q) = \|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2 - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= \langle \mu_P, \mu_P\rangle_{\mathcal{H}} + \langle \mu_Q, \mu_Q\rangle_{\mathcal{H}} - 2\langle \mu_P, \mu_Q\rangle_{\mathcal{H}}$$
$$= E_{X,X'}\left[\langle k_X, k_{X'}\rangle_{\mathcal{H}}\right] + \mathbb{E}_{Y,Y'}\left[\langle k_Y, k_{Y'}\rangle_{\mathcal{H}}\right] - 2\mathbb{E}_{X,Y}\left[\langle k_X, k_Y\rangle_{\mathcal{H}}\right]$$

Second plug-in estimator of the MMD

$$\widehat{MMD}^2_{u,k}(P,Q)$$
$$= \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} k(X_i,X_j) + \frac{1}{m(m-1)}\sum_{p\neq q=1}^{m} k(Y_p,Y_q)$$
$$-2\frac{\sum_{i=1}^n\sum_{j=1}^m k(X_i,Y_j)}{mn}$$

Theorem

$$\mathbb{E}\left[\widehat{MMD}^2_{u,k}(P,Q)\right] = MMD_k^2(P,Q)$$

Remark: Is there another choice for an unbiased estimator?

Testing for two fixed populations

Two-sample test

$$X_1, \ldots, X_n \sim P \text{ and } Y_1, \ldots, Y_m \sim Q, \text{ with } P = Q?$$

Statistical test

- Hypothesis:

$$\mathrm{H}_0 : P = Q \text{ (no change)} \quad \textbf{vs} \quad \mathrm{H}_1 : P \neq Q$$

- Rejection region:

$$\mathcal{R} = \left\{ \widehat{MMD}_k^2(P, Q) > \eta_\alpha \right\}$$

where $\eta_\alpha > 0$ depends on the Type-I error

Problem

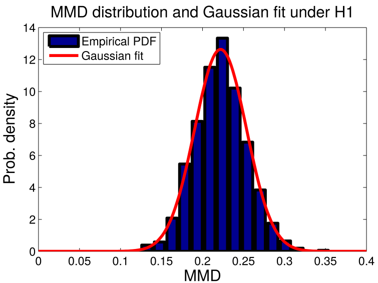
Distribution of $\widehat{MMD}_k^2(P, Q)$ difficult to estimate

Distribution of the unbiased estimator

$$\widehat{MMD}^2_{u,k}(P,Q)$$
$$= \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} k(X_i, X_j) + \frac{1}{m(m-1)} \sum_{p \neq q = 1}^{m} k(Y_p, Y_q)$$
$$- 2 \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} k(X_i, Y_j)}{mn}$$

If $P \neq Q$ $(m = n)$

$\widehat{MMD}^2_{u,k}(P,Q)$: Gaussian asymptotic distrib. (U-stat.)



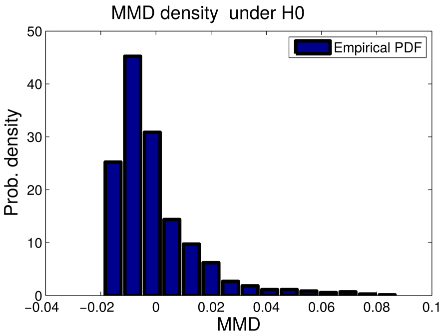
Distribution of the unbiased estimator

If $P = Q$ $(m = n)$

$$m\widehat{MMD}_k^2(P,Q) \sim 2\sum_{\ell=1}^{+\infty} \lambda_\ell \left[\chi_\ell^2 - 1 \right]$$

where, for all $\ell \geq 1$,

$$\int_{\mathcal{X}} k_x \psi_\ell(x) dP(x) = \lambda_\ell \psi_\ell \in \mathcal{H}$$



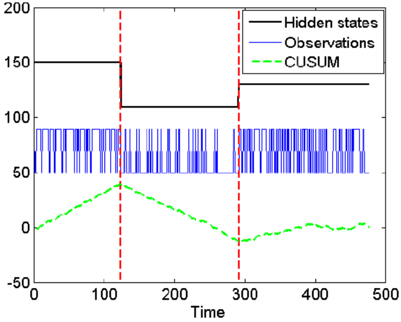
Sequential testing for one change-point

Single changepoint detection

- Time series: $(t_1, X_1), \ldots, (t_n, X_n) \in [0, +\infty] \times \mathbb{R}$
- **Assumption:**

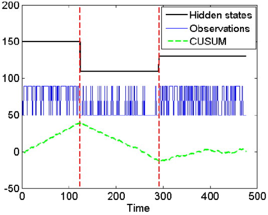
$$P_{X_1} = \cdots = P_{X_r} \neq P_{X_{r+1}} = \cdots = P_{X_n}$$

- Usually replaced by the mean (distributional features)

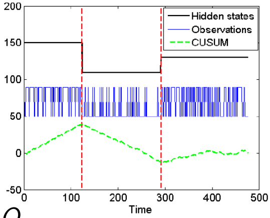


From two-sample test to Single changepoint detection

- r: changepoint location
- $P_{X_r} \neq P_{X_{r+1}}$: "abrupt" change

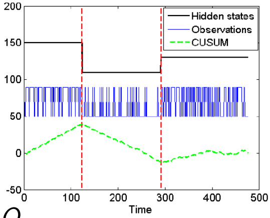


From two-sample test to Single changepoint detection



- r: changepoint location
- $P_{X_r} \neq P_{X_{r+1}}$: "abrupt" change
- Two populations:
  - From 1 to r: all X_is from P
  - From $r+1$ to n: all X_is from Q
- Each time $1 \leq t \leq n$: candidate changepoint

From two-sample test to Single changepoint detection



- r: changepoint location
- $P_{X_r} \neq P_{X_{r+1}}$: "abrupt" change
- Two populations:
  - From 1 to r: all X_is from P
  - From $r+1$ to n: all X_is from Q
- Each time $1 \leq t \leq n$: candidate changepoint

MMD-based statistic

$$\widehat{MMD}^2_{u,k}(t) = \|\widehat{\mu}_P(1:t) - \widehat{\mu}_Q(t+1:n)\|^2_{\mathcal{H}}$$
$$= \frac{\sum_{i\neq j=1}^{t} k(X_i, X_j)}{t(t-1)} + \frac{\sum_{p\neq q=t+1}^{n} k(Y_p, Y_q)}{(n-t)(n-t-1)}$$
$$- 2\frac{\sum_{i=1}^{t}\sum_{j=t+1}^{n} k(X_i, Y_j)}{t(n-t)}$$

Single changepoint detection algorithm

Algorithm

1. Compute

$$\mathcal{M}_n = \min_{1 \leq t \leq n} \left\{ \widehat{MMD}^2_{u,k}(t) \right\}$$

Single changepoint detection algorithm

Algorithm

1. Compute

$$\mathcal{M}_n = \min_{1\leq t\leq n} \left\{ \widehat{MMD}^2_{u,k}(t) \right\}$$

2. Rejection region:

$$\mathcal{R}_\alpha = \{\mathcal{M}_n \geq q_\alpha\}$$

q_α: $\alpha-$quantile of $P_{\mathcal{M}_n}$ under the null

Single changepoint detection algorithm

Algorithm

1. Compute

$$\mathcal{M}_n = \min_{1 \leq t \leq n} \left\{ \widehat{MMD}^2_{u,k}(t) \right\}$$

2. Rejection region:

$$\mathcal{R}_\alpha = \{\mathcal{M}_n \geq q_\alpha\}$$

   q_α: $\alpha-$quantile of $P_{\mathcal{M}_n}$ under the null

3. If the null hypothesis is rejected, then

$$\hat{r}_\alpha = Arg \min_{1 \leq t \leq n} \left\{ \widehat{MMD}^2_{u,k}(t) \right\}$$

Single changepoint detection algorithm

Algorithm

1. Compute

$$\mathcal{M}_n = \min_{1 \le t \le n} \left\{ \widehat{MMD}^2_{u,k}(t) \right\}$$

2. Rejection region:

$$\mathcal{R}_\alpha = \{ \mathcal{M}_n \ge q_\alpha \}$$

   q_α: $\alpha-$quantile of $P_{\mathcal{M}_n}$ under the null

3. If the null hypothesis is rejected, then

$$\hat{r}_\alpha = Arg \min_{1 \le t \le n} \left\{ \widehat{MMD}^2_{u,k}(t) \right\}$$

Remarks:

- q_α computed by bootstrap or approximation (asymptotic Gaussian)
- Sequential approaches also possible (multiple changepoints)

Independent random variables?

Two samples

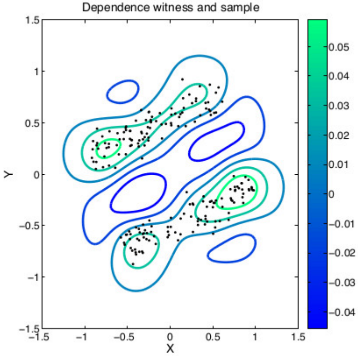
- $(X_i, Y_i)_{i=1}^{n}$: n couples
- $X_1, \ldots, X_n \sim P_X$: realizations of X
- $Y_1, \ldots, Y_n \sim P_Y$: realizations of Y

Independent random variables?

Two samples

- $(X_i, Y_i)_{i=1}^n$: n couples
- $X_1, \ldots, X_n \sim P_X$: realizations of X
- $Y_1, \ldots, Y_n \sim P_Y$: realizations of Y

Are X and Y independent random variables?



Testing for independence

Statistical test

- Null hypothesis:
$$\mathrm{H}_0 : P_{X,Y} = P_X \otimes P_Y$$
- Alternative:
$$\mathrm{H}_1 : P_{X,Y} \neq P_X \otimes P_Y$$

Testing for independence

Statistical test

- Null hypothesis:

$$H_0 : P_{X,Y} = P_X \otimes P_Y$$

- Alternative:

$$H_1 : P_{X,Y} \neq P_X \otimes P_Y$$

Idea

- $P = P_{X,Y}$
- $Q = P_X \otimes P_Y$

Testing for independence

Statistical test

- Null hypothesis:

$$\mathrm{H}_0 : P_{X,Y} = P_X \otimes P_Y$$

- Alternative:

$$\mathrm{H}_1 : P_{X,Y} \neq P_X \otimes P_Y$$

Idea

- $P = P_{X,Y}$
- $Q = P_X \otimes P_Y$

$$MMD_k(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}_X,\mathcal{H}_Y}$$

HSIC criterion for independence testing

Product kernel on $\mathcal{X} \times \mathcal{Y}$

$$k((X_i, Y_i), (X_j, Y_j)) = k_X(X_i, X_j) \times k_Y(Y_i, Y_j)$$

- $\mathcal{H}_X$: RKHS of k_X
- $\mathcal{H}_Y$: RKHS of k_Y

HSIC criterion for independence testing

Product kernel on $\mathcal{X} \times \mathcal{Y}$

$$k((X_i, Y_i), (X_j, Y_j)) = k_X(X_i, X_j) \times k_Y(Y_i, Y_j)$$

- $\mathcal{H}_X$: RKHS of k_X
- $\mathcal{H}_Y$: RKHS of k_Y
- Then: $k(\cdot, \cdot)$: reproducing kernel on $\mathcal{H}_X \times \mathcal{H}_Y$

Remark:

Other choices of kernels are possible

HSIC

$$\begin{aligned} HSIC(P_{X,Y}, P_X \otimes P_Y) &= \left\| \mu_{P_{X,Y}} - \mu_{P_X \otimes P_Y} \right\|^2_{\mathcal{H}_X, \mathcal{H}_Y} \\ &= MMD_k^2(P_{X,Y}, P_X \otimes P_Y) \end{aligned}$$

HSIC criterion for independence testing

Product kernel on $\mathcal{X} \times \mathcal{Y}$

$$k((X_i, Y_i), (X_j, Y_j)) = k_X(X_i, X_j) \times k_Y(Y_i, Y_j)$$

- $\mathcal{H}_X$: RKHS of k_X
- $\mathcal{H}_Y$: RKHS of k_Y
- Then: $k(\cdot, \cdot)$: reproducing kernel on $\mathcal{H}_X \times \mathcal{H}_Y$

Remark:

Other choices of kernels are possible

HSIC

$$\begin{aligned} HSIC(P_{X,Y}, P_X \otimes P_Y) &= \left\| \mu_{P_{X,Y}} - \mu_{P_X \otimes P_Y} \right\|^2_{\mathcal{H}_X, \mathcal{H}_Y} \\ &= MMD_k^2(P_{X,Y}, P_X \otimes P_Y) \end{aligned}$$

<w**Exercise:** Compute $\left\| \mu_{P_{X,Y}} \right\|^2_{\mathcal{H}_X \times \mathcal{H}_Y}$ and $\left\| \mu_{P_X \otimes P_Y} \right\|^2_{\mathcal{H}_X \times \mathcal{H}_Y}$