Comparison of Empirical Probability Distributions

Sylvain Combettes

Supervisors: Alexandre Voisin, Pierre Vallois

February 14, 2020

MINES NANCY ARTEM

UNIVERSITÉ DE LORRAINE

Introduction

Why compare data sets' distributions

- Why? Important problem in modelling.
- Example: maintenance / anomaly detection
  $\rightarrow$ detect when a distribution is "shifted" from its "normal" state
- 2 main categories:
  - integral probability metrics (IPMs)
  - f-divergences
- Application: Choquet integral with stochastic inputs

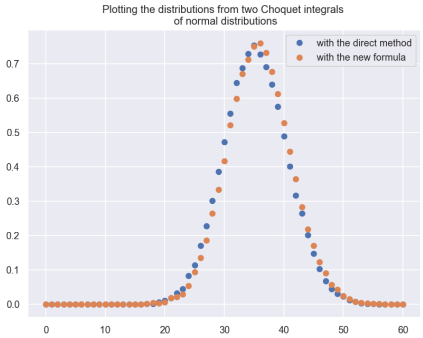


Table of contents

I – Integral Probability Metrics (IPMs)

Integral probability metric γ

IPMs: empirical estimation of distances on probabilities on $S \subset \mathbb{R}$.

Definition (Integral probability metric γ)

Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on a measurable set $S \subset \mathbb{R}$, the **integral probability metric** (IPM) giving the distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_S f \, \mathrm{d}\mathbb{P} - \int_S f \, \mathrm{d}\mathbb{Q} \right| \tag{1}$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on S.

Each choice of $\mathcal{F}$ leads to a specific IPM.

Focus on one IPM: Kantorovich metric W.

Kantorovich metric W

Definition (Kantorovich metric W)

Setting:

$$\boxed{\mathcal{F} = \{f : \|f\|_L \leqslant 1\}} \tag{2}$$

in (1) yields the **Kantorovich metric** W, where $\|f\|_L$ is the Lipschitz semi-norm of a bounded continuous real-valued function f:

$$\|f\|_L = \sup\left\{\frac{|f(x) - f(y)|}{|x - y|} : x \neq y \text{ in } S \subset \mathbb{R}\right\} \tag{3}$$

Notation: $\mathcal{F}_W = \{f : \|f\|_L \leqslant 1\}$.

Example: explicit computation of W

- Let $S = [a, s]$ and $h = 1$ (interval length).
- Suppose $\mathbb{P} = \mathbb{U}([a, a + h])$ and $\mathbb{Q} = \mathbb{U}([r, r + h])$, where:

$$-\infty < a \leqslant r \leqslant a + h \leqslant r + h < \infty \tag{4}$$

- Then, we can show that:

$$\boxed{W(\mathbb{P}, \mathbb{Q}) = r - a} \tag{5}$$

- W depends on the parameters a and r of the uniform distributions:
  - $r - a \nearrow \Longrightarrow W(\mathbb{P}, \mathbb{Q}) \nearrow$
  - $W(\mathbb{P}, \mathbb{Q}) = 0 \Longleftrightarrow r = a \Longleftrightarrow \mathbb{P} = \mathbb{Q}$

Definition (Empirical estimator of the Kantorovich metric)

Given $\left\{X_1^{(1)}, X_2^{(1)}, \ldots, X_m^{(1)}\right\}$ and $\left\{X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}\right\}$, which are i.i.d. samples drawn randomly from $\mathbb{P}$ and $\mathbb{Q}$, respectively, the **empirical estimator of $W(\mathbb{P}, \mathbb{Q})$** is:

$$W\left(\mathbb{P}_m, \mathbb{Q}_n\right) = \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f\left(X_i^{(1)}\right) - \frac{1}{n} \sum_{j=1}^{n} f\left(X_i^{(2)}\right) \right| \tag{6}$$

where $\mathbb{P}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i^{(1)}}$ and $\mathbb{Q}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i^{(2)}}$ represent the empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$, respectively, and $N = n + m$.

Goal: find the function f that solves (6) for $\mathcal{F} = \mathcal{F}_W$.

Theorem (Empirical estimator of the Kantorovich metric)

We have:

$$W\left(\mathbb{P}_m, \mathbb{Q}_n\right) = \sum_{i=1}^{N} \widetilde{Y}_i a_i^\star \tag{7}$$

$$\widetilde{Y}_i = \frac{1}{m} \text{ when } X_i = X_i^{(1)} \text{ for } i = 1, \ldots, m \tag{8}$$

$$\widetilde{Y}_{m+i} = -\frac{1}{n} \text{ when } X_{m+i} = X_i^{(2)} \text{ for } i = 1, \ldots, n$$

and $\left\{a_i^\star\right\}_{i=1}^{N}$ *solve the following linear program:*

$$\max_{a_1,\ldots,a_N} \left\{ \sum_{i=1}^{N} \widetilde{Y}_i a_i : -|X_i - X_j| \leqslant a_i - a_j \leqslant |X_i - X_j|, \forall i, j \right\} \tag{9}$$

➥ In practice: `PuLP` library from Python.

Solving the linear programming problem for $N = 4$

The objective function is: $\sum_{i=1}^{N} \widetilde{Y}_i a_i = \widetilde{Y}_1 a_1 + \widetilde{Y}_2 a_2 + \widetilde{Y}_3 a_3 + \widetilde{Y}_4 a_4$

The constraints are:

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} \leqslant \begin{pmatrix} |X_1 - X_2| \\ |X_1 - X_2| \\ |X_1 - X_3| \\ |X_1 - X_3| \\ |X_1 - X_4| \\ |X_1 - X_4| \\ |X_2 - X_3| \\ |X_2 - X_3| \\ |X_2 - X_4| \\ |X_2 - X_4| \\ |X_3 - X_4| \\ |X_3 - X_4| \end{pmatrix} \tag{10}$$

Memory issue: $p = N(N-1)$, e.g. $N = 200 \rightarrow p \times N = 7\,960\,000$.

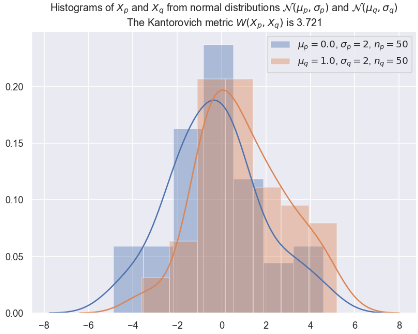
How does the Kantorovich metric W evolve?

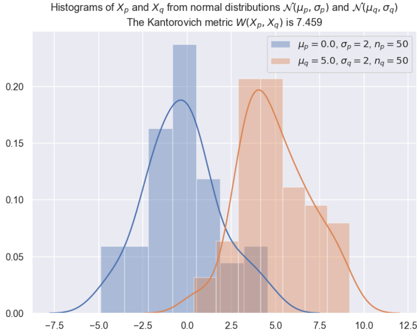
- running several simulations using Python
- samples in 1D from two normal distributions $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$
- X_p are the n_p samples drawn from $\mathbb{P}$
  X_q the n_q samples drawn from $\mathbb{Q}$
  - How does $W(X_p, X_q)$ evolve with $\mu_q - \mu_p$?
  - How does $W(X_p, X_q)$ evolve with $\sigma_q - \sigma_p$?
  - How does $W(X_p, X_q)$ evolve with $n_q - n_p$?
- IPMs input empirical samples ($N \leqslant 1\,000$)
- assessing a linear regression model with R^2 using `scikit-learn` (Python)

Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

Influence of $\mu_q - \mu_p$

Histograms and approx. density of the samples X_p and X_q:

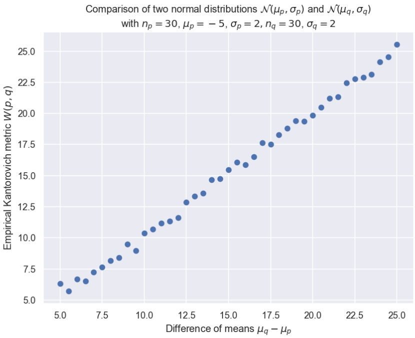


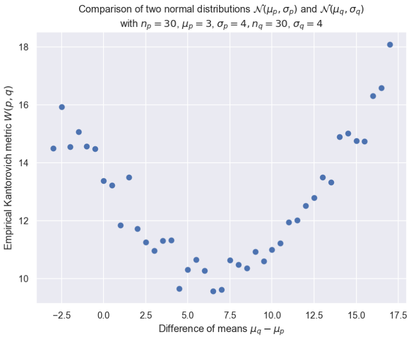


Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

Influence of $\mu_q - \mu_p$

Evolution of $W(X_p, X_q)$ with $\mu_q - \mu_p$:

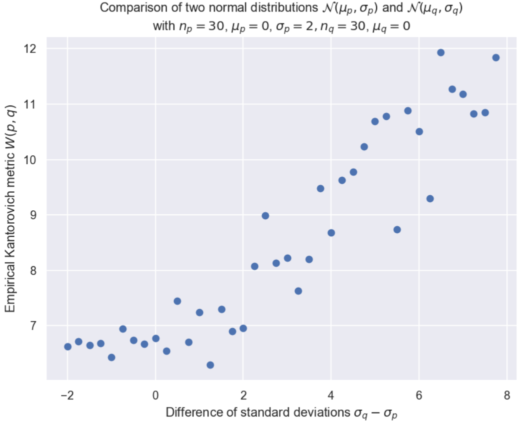




Is the dependency of $W(X_p, X_q)$ to $\mu_q - \mu_p$ linear? $R^2 = 0.997$.

Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

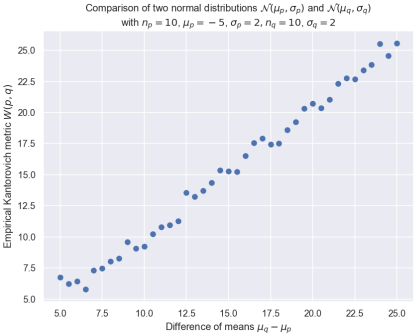
Influence of $\sigma_q - \sigma_p$

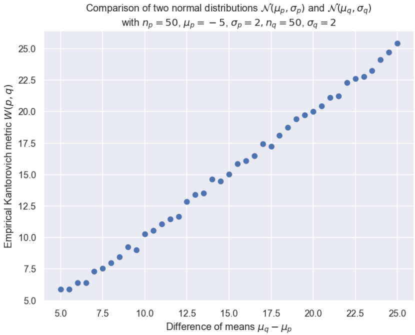


Is the dependency of $W(X_p, X_q)$ to $\sigma_q - \sigma_p$ linear? $R^2 = 0.856$.

Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

Influence of $n_p = n_q$





We will not consider the number of samples as a relevant parameter:

- $n_p = n_q = 10 \implies R^2 = 0.991$
- $n_p = n_q = 30 \implies R^2 = 0.997$
- $n_p = n_q = 50 \implies R^2 = 0.998$

II – f-divergences

A f-divergence is a function $D_f(\mathbb{P}, \mathbb{Q})$ that measures the difference between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$.

We will focus on $S \subset \mathbb{R}$.

Definition (f-divergence D_f (discrete version))

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a measurable set $S \subset \mathbb{R}$. Let f be a continuous convex real function on $\mathbb{R}_+$, with $f(1) = 0$. Then, the ***f*-divergence of $\mathbb{P}$ from $\mathbb{Q}$** is defined as:

$$\boxed{D_f(\mathbb{P}, \mathbb{Q}) = \sum_{x \in S} \mathbb{Q}(x)\, f\left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\right)} \tag{11}$$

Each choice of f in (11) leads to a particular f-divergence.

Kullback-Leibler divergence

Definition

We choose $f(u) = u \log(u)$ in (11).

Definition (Kullback-Leibler divergence D_{KL})

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a measurable set $S \subset \mathbb{R}$. The **Kullback-Leibler divergence** (or KL divergence) of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:

$$\boxed{D_{\text{KL}}(\mathbb{P}, \mathbb{Q}) = \sum_{x \in S} \mathbb{P}(x) \log \left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)} \right)} \tag{12}$$

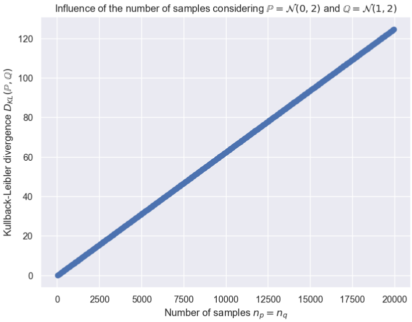
D_{KL} is non-negative and is 0 if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution.

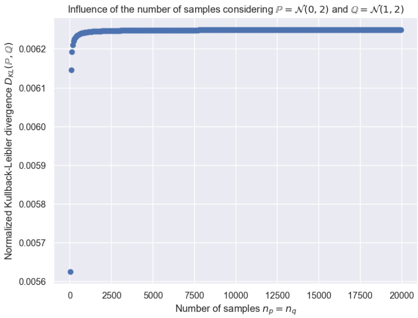
It is not a true distance because it is not symmetric: $D_{\text{KL}}(\mathbb{P}, \mathbb{Q}) \neq D_{\text{KL}}(\mathbb{Q}, \mathbb{P})$ for some $\mathbb{P}$ and $\mathbb{Q}$.

Kullback-Leibler divergence

Influence of the number of samples

Two (discrete) normal distributions $\mathbb{P} = \mathcal{N}(0, 2)$ and $\mathbb{Q} = \mathcal{N}(1, 2)$.





➥ The KL divergence needs to be "normalized".

Hellinger distance

We choose $f(u) = (\sqrt{u} - 1)^2$ in (11).

Definition (Hellinger distance D_{H})

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a measurable set $S \subset \mathbb{R}$. The **Hellinger distance** of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:

$$D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) = \sum_{x \in S} \left(\sqrt{\mathbb{P}(x)} - \sqrt{\mathbb{Q}(x)} \right)^2 \tag{13}$$

D_{H} is non-negative, is 0 if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution and is symmetric.

D_{H} is a true distance.

D_{H} needs to be "normalized".

Variational distance

We choose $f(u) = |u - 1|$ in (11).

Definition (Variational distance D_V)

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a measurable set $S \subset \mathbb{R}$. The **Variational distance** of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:

$$D_V(\mathbb{P}, \mathbb{Q}) = \sum_{x \in S} |\mathbb{P}(x) - \mathbb{Q}(x)| \tag{14}$$

D_V is non-negative, is 0 if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution and is symmetric.

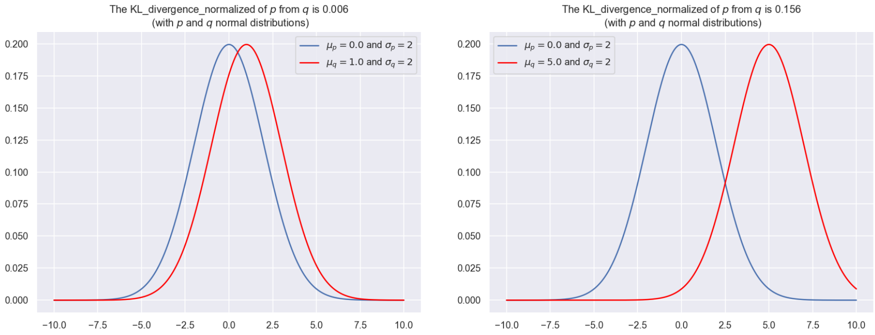
D_V is a true distance.

D_V needs to be "normalized".

Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

Influence of $\mu_q - \mu_p$

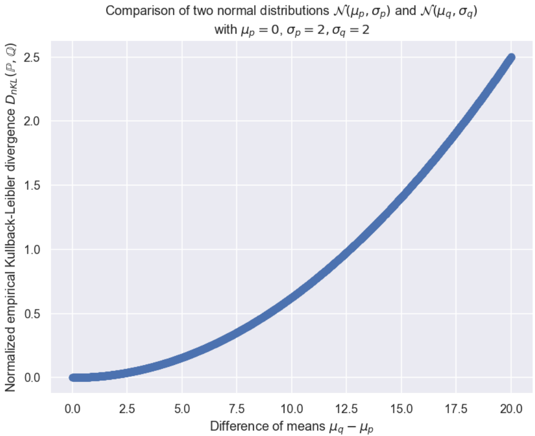
Two (discrete) normal distributions $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$.



Note: $n_p = n_q = 20\ 000$.

Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

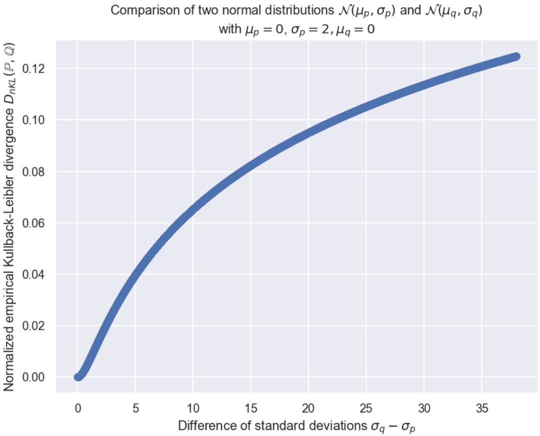
Influence of $\mu_q - \mu_p$



Is the dependency of $D_{\text{nKL}}(\mathbb{P}, \mathbb{Q})$ to $(\mu_q - \mu_p)^2$ linear? $R^2 = 1$.

Comparison of $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$

Influence of $\sigma_q - \sigma_p$



Is the dependency of $D_{\mathrm{nKL}}(\mathbb{P}, \mathbb{Q})$ to $\sqrt{(\sigma_q - \sigma_p)}$ linear? $R^2 = 0.991$.

III – Application to the Choquet integral

Informal definition

The Choquet integral is a non-linear aggregation operator.

Finite set $S = \{1, 2, \dots, n\}$. S is a set of criteria.

Definition (Choquet integral C of a vector G)

Let $G = (G_1, \dots, G_n) \in \mathbb{R}^n$. The **Choquet integral of G with respect to K** is the real number:

$$\boxed{C_K(G) = \sum_{i=1}^{n} G_{\sigma(i)} K_{\sigma(i)}} \tag{15}$$

with σ a permutation of the values of S such that $G_{\sigma(1)} \leqslant \dots \leqslant G_{\sigma(n)}$ and $K \in \mathbb{R}^n$.

We will consider stochastic entries: $G = (G_1, \dots, G_n)$ with $G_i \hookrightarrow \mathcal{N}(\mu, \sigma)$. $C_K(G)$ is a random variable.

Context: comparing two methods for computing a Choquet integral

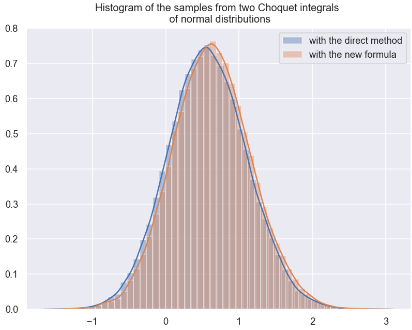
- synthetic samples and not real industrial ones
- computing the Choquet integral with stochastic entries
  - 1st method $\rightarrow$ direct method : Monte-Carlo simulation
    $\rightarrow$ X_p and $\mathbb{P}$
  - 2nd method $\rightarrow$ new formula giving the distribution of the values taken by the Choquet integral
    $\rightarrow$ $\mathbb{Q}$ and X_q (drawn from $\mathbb{Q}$)
- goal: verify experimentally that the new formula gives "acceptable" results
- in practice: compare the distance between the distributions from the 2 methods: $\mathbb{P}$ and $\mathbb{Q}$

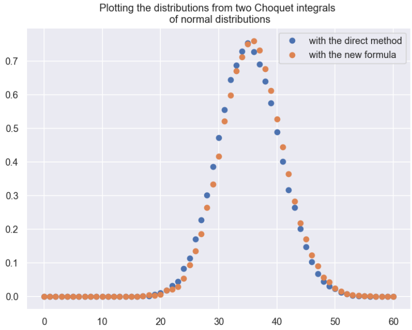
Presenting the data

Choquet integral of normal distributions.

IPMs input the samples X_p and X_q.

f-divergences input the empirical distributions $\mathbb{P}$ and $\mathbb{Q}$.





IPMs and f-divergences

Empirical results:

Average Kantorovich metric W	0.894 ± 0.033
Normalized Kullback-Leibler divergence D_{nKL}	Inf
Normalized Hellinger distance D_{nH}	5.179×10^{-4}
Normalized Variational distance D_{nV}	1.381×10^{-2}

➥ These values are "very small"

- → the distance between $\mathbb{P}$ and $\mathbb{Q}$ is "very small"
- → $\mathbb{P}$ and $\mathbb{Q}$ are "very close"
- → the two methods give "very similar" results
- → the new formula is "correct"

Conclusion

Conclusion

Let $\boxed{\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)}$ and $\boxed{\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)}$. X_p drawn from $\mathbb{P}$ and X_q from $\mathbb{Q}$.

- integral probability metrics (IPMs)
  - each choice of $\mathcal{F}$ leads to a specific IPM
  - focus on the Kantorovich metric W
    - need to solve a linear programming problem $\rightarrow$ memory issue because $p = N(N-1) \rightarrow N \leqslant 1\,000$
    - $\mu_q - \mu_p \nearrow \implies W(X_p, X_q) \nearrow$
    - $\sigma_q - \sigma_p \nearrow \implies W(X_p, X_q) \nearrow$
- f-divergences
  - each choice of f leads to a specific f-divergence
  - Kullback-Leibler divergence D_{KL} (not symmetric)
    - $\mu_q - \mu_p \nearrow \implies D_{\mathrm{KL}}(\mathbb{P}, \mathbb{Q}) \nearrow$
    - $\sigma_q - \sigma_p \nearrow \implies D_{\mathrm{KL}}(\mathbb{P}, \mathbb{Q}) \nearrow$
  - need to "normalize"
- Application to the Choquet integral $\rightarrow$ the new formula gives "similar" results to the direct method

Thanks for listening.