Structured Data: Dependency, Testing

Zoltán Szabó

$\in$ Structured Data: Learning, Prediction, Dependency, Testing
M2 Data Science, University of Paris-Saclay

Paris, France
March 5, 12, 19, 26, 2018

Contact information & to-know

- Email:

  zoltan (dot) szabo (at) polytechnique (dot) edu
- Web:

  http://www.cmap.polytechnique.fr/~zoltan.szabo/

Contact information & to-know

- Email:

  zoltan (dot) szabo (at) polytechnique (dot) edu

- Web:

  http://www.cmap.polytechnique.fr/~zoltan.szabo/

- Deadlines (if you choose from Stack-2): send me by
  - March 12: the selected paper.
  - March 26: the critical analysis.

I will make the candidate papers $\in$ Stack-2 available with the slides.

Software (Python, Matlab)

- Dependency measures (KCCA, HSIC), divergences (MMD), etc.; several demos:

  https://bitbucket.org/szzoli/ite-in-python
  https://bitbucket.org/szzoli/ite/

- 2-sample, independence & goodness-of-fit tests (quadratic $\rightarrow$ linear-time methods):

  https://github.com/wittawatj/interpretable-test
  https://github.com/wittawatj/fsic-test
  https://github.com/wittawatj/kernel-gof

Outline

- Motivation:
  - Objective functions: from dependency measures.
  - Testing.

Outline

- Motivation:
  - Objective functions: from dependency measures.
  - Testing.
- Kernel, RKHS.
- Kernel canonical correlation analysis.
- Mean embedding:
  - Characteristic property,
  - Universality.
- Maximum mean discrepancy.
- Cross-covariance operator, HSIC.
- Hypothesis testing.

Dependency Measures as Objective Functions

Outlier-robust image registration [Kybic, 2004, Neemuchwala et al., 2007]

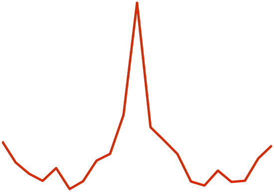
Given two images:

Goal: find the transformation which takes the right one to the left.

Outlier-robust image registration [Kybic, 2004, Neemuchwala et al., 2007]

Given two images:





Goal: find the transformation which takes the right one to the left.

Outlier-robust image registration: equations

- Reference image: $\mathbf{y}_{\text{ref}}$,
- test image: $\mathbf{y}_{\text{test}}$,
- possible transformations: $\boldsymbol{\Theta}$.

Objective:

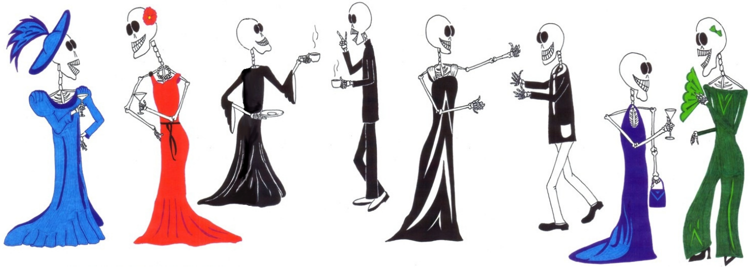
$$J(\boldsymbol{\theta}) = \underbrace{I(\mathbf{y}_{\text{ref}}, \mathbf{y}_{\text{test}}(\boldsymbol{\theta}))}_{\text{similarity}} \to \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}},$$

In the example: I=KCCA.

Independent Subspace Analysis [Cardoso, 1998]

Cocktail party problem:

- independent groups of people / music bands,
- observation = mixed sources.



ISA equations

Observation:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right].$$

Goal: $\hat{\mathbf{s}}$ from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Assumptions:

- independent groups: $I\left(\mathbf{s}^1, \ldots, \mathbf{s}^M\right) = 0$,
- $\mathbf{s}^m$-s: non-Gaussian,
- $\mathbf{A}$: invertible.

ISA solution

Find **W** which makes the estimated components independent:

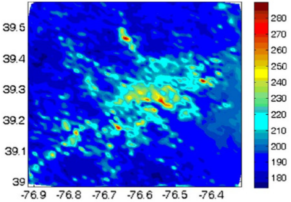
$$\mathbf{y} = \mathbf{W}\mathbf{x} = \left[\mathbf{y}^1; \ldots; \mathbf{y}^M\right],$$

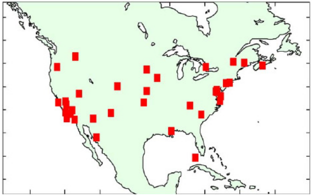
$$J(\mathbf{W}) = I\left(\mathbf{y}^1, \ldots, \mathbf{y}^M\right) \to \min_{\mathbf{W}}.$$

Distribution regression [Póczos et al., 2013, Szabó et al., 2016]. Sustainability

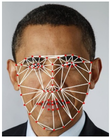
- **Goal**: aerosol prediction = air pollution $\rightarrow$ climate.

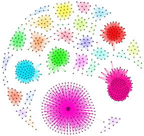
- Prediction using labelled bags:
  - bag := multi-spectral satellite measurements over an area,
  - label := local aerosol value.





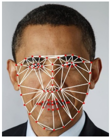
Objects in the bags

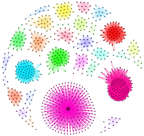




- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, . . .

Objects in the bags





- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, ...
- Wider context (statistics): point estimation tasks.

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} \Big[f\underbrace{\left(\mu_{\hat{P}_i}\right)}_{\text{feature of } \hat{P}_i} - y_i \Big]^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2 .$$

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\mu_{\hat{P}_i}\right) - y_i\right]^2 + \lambda \|f\|_{\mathcal{H}}^2 .$$

- Prediction:

$$\hat{y}\left(\hat{P}\right) = \mathbf{g}^T (\mathbf{G} + \ell\lambda\mathbf{I})^{-1}\mathbf{y},$$

$$\mathbf{g} = \left[K\left(\mu_{\hat{P}}, \mu_{\hat{P}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right)\right], \mathbf{y} = [y_i].$$

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \underset{f \in \mathcal{H}_K}{\arg\min} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\mu_{\hat{P}_i}\right) - y_i \right]^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2 .$$

- Prediction:

$$\hat{y}\left(\hat{P}\right) = \mathbf{g}^T (\mathbf{G} + \ell\lambda \mathbf{I})^{-1} \mathbf{y},$$
$$\mathbf{g} = \left[K\left(\mu_{\hat{P}}, \mu_{\hat{P}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right)\right], \mathbf{y} = [y_i].$$

Challenge

Inner product of distributions: $K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right) = ?$

Feature selection

- **Goal**: find
  - the feature subset (# of rooms, criminal rate, local taxes)
  - most relevant for house price prediction (y).

Feature selection: equations

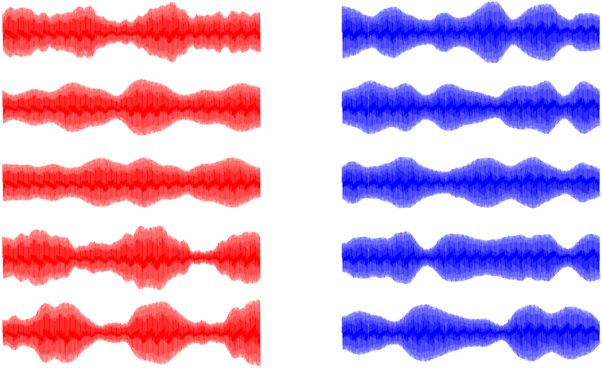
- Features: $x^1, \ldots, x^F$. Subset: $S \subseteq \{1, \ldots, F\}$.
- MaxRelevance - MinRedundancy principle [Peng et al., 2005]:

$$J(S) = \frac{1}{|S|} \sum_{i \in S} I\left(x^i, y\right) - \frac{1}{|S|^2} \sum_{i,j \in S} I\left(x^i, x^j\right) \to \max_{S \subseteq \{1,\ldots,F\}}.$$

Testing

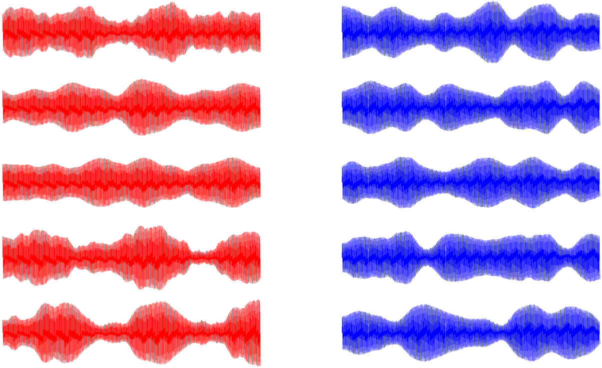
Motivation: detecting differences in AM signals

- Amplitude modulation:
  - simple technique to transmit voice over radio.
  - in the example: 2 songs.
- Fragments from $\text{song}_1 \sim \mathbb{P}_x$, $\text{song}_2 \sim \mathbb{P}_y$.



Motivation: detecting differences in AM signals

- Amplitude modulation:
  - simple technique to transmit voice over radio.
  - in the example: 2 songs.
- Fragments from $\text{song}_1 \sim \mathbb{P}_x$, $\text{song}_2 \sim \mathbb{P}_y$.



Question: $\mathbb{P}_x = \mathbb{P}_y$?

Motivation: discrete domain - 2-sample testing

- How do we compare distributions?
- Given: 2 sets of text fragments (fisheries, agriculture).

x_1: Now disturbing reports out of Newfoundland show that the fragile snow crab industry is in serious decline. First the west coast salmon, the east coast salmon and the cod, and now the snow crabs off Newfoundland.

x_2: To my pleasant surprise he responded that he had personally visited those wharves and that he had already announced money to fix them. What wharves did the minister visit in my riding and how much additional funding is he going to provide for Delaps Cove, Hampton, Port Lorne, . . .

. . .

y_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

y_2: On the grain transportation system we have had the Estey report and the Kroeger report. We could go on and on. Recently programs have been announced over and over by the government such as money for the disaster in agriculture on the prairies and across Canada.

. . .

Motivation: discrete domain - 2-sample testing

- How do we compare distributions?
- Given: 2 sets of text fragments (fisheries, agriculture).

x_1: Now disturbing reports out of Newfoundland show that the fragile snow crab industry is in serious decline. First the west coast salmon, the east coast salmon and the cod, and now the snow crabs off Newfoundland.

x_2: To my pleasant surprise he responded that he had personally visited those wharves and that he had already announced money to fix them. What wharves did the minister visit in my riding and how much additional funding is he going to provide for Delaps Cove, Hampton, Port Lorne, . . .

. . .

y_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

y_2: On the grain transportation system we have had the Estey report and the Kroeger report. We could go on and on. Recently programs have been announced over and over by the government such as money for the disaster in agriculture on the prairies and across Canada.

. . .

Do $\{x_i\}$ and $\{y_j\}$ come from the same distribution, i.e. $\mathbb{P}_x = \mathbb{P}_y$?

Motivation: discrete domain - independence testing

- How do we detect dependency? (paired samples)

x_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

x_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

...

y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reu de cet argent.

y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

...

Motivation: discrete domain - independence testing

- How do we detect dependency? (paired samples)

x_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

x_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

...

y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reu de cet argent.

y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

...

Are the French paragraphs translations of the English ones, or have nothing to do with it, i.e. $\mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y$?

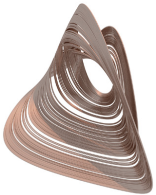
We will use kernels to tackle these problems

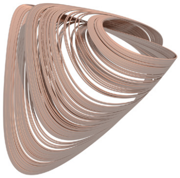
They exist essentially on any data type

We will use kernels to tackle these problems

They exist essentially on any data type

trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002], time series [Cuturi, 2011], strings [Lodhi et al., 2002], mixture models, hidden Markov models or linear dynamical systems [Jebara et al., 2004], sets [Haussler, 1999, Gärtner et al., 2002], fuzzy domains [Guevara et al., 2017], distributions [Hein and Bousquet, 2005, Martins et al., 2009, Muandet et al., 2011], groups [Cuturi et al., 2005] with specific constructions on permutations [Jiao and Vert, 2016], graphs [Vishwanathan et al., 2010, Kondor and Pan, 2016], ...





Kernel Canonical Correlation Analysis (KCCA)

Independence measures

- Given: random variable $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $(x, y) \sim \mathbb{P}_{xy}$.
- Goal: measure the dependence of x and y.

Independence measures

- Given: random variable $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $(x, y) \sim \mathbb{P}_{xy}$.
- Goal: measure the dependence of x and y.
- Desiderata for a $Q(\mathbb{P}_{xy})$ independence measure [Rényi, 1959]:
    1. $Q(\mathbb{P}_{xy})$ is well-defined,
    2. $Q(\mathbb{P}_{xy}) \in [0, 1]$,
    3. $Q(\mathbb{P}_{xy}) = 0$ iff. $x \perp y$.
    4. $Q(\mathbb{P}_{xy}) = 1$ iff. $y = f(x)$ or $x = g(y)$.

- He showed:

$$Q(\mathbb{P}_{xy}) = \sup_{f,g:\ \text{measurable}} \operatorname{corr}(f(x), g(y)),$$

satisfies 1-4.

- Too ambitious:
  - computationally intractable.
  - many measurable functions.

Independence measures: measurable $\to$ continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X}$ metric $\to \mathbb{R}$, bounded continuous$\}$ would also work.
- Still too large!

Independence measures: measurable → continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \text{ metric} \to \mathbb{R}, \text{ bounded continuous}\}$ would also work.
- Still too large!
- Idea:
  - certain RKHS-s are dense in $C_b(\mathcal{X})$.
  - computionally tractable.

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated:
  - feature maps $\varphi(x) = k(\cdot, x)$, $\psi(y) = \ell(\cdot, y)$,
  - RKHS-s $\mathcal{H}_k$, $\mathcal{H}_\ell$.

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated:
  - feature maps $\varphi(x) = k(\cdot, x)$, $\psi(y) = \ell(\cdot, y)$,
  - RKHS-s $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- KCCA measure of $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)),$$

$$\operatorname{corr}(f(x), g(y)) = \frac{\operatorname{cov}_{xy}(f(x), g(y))}{\sqrt{\operatorname{var}_x f(x) \operatorname{var}_y g(y)}}.$$

KCCA: notes

- Optimization domain: $\mathcal{H}_k \times \mathcal{H}_\ell \ni (f, g)$.
- By reproducing property: we will get a finite-D task.
- k,ℓ linear: traditional CCA.
- In practice: we have $\{(x_n, y_n)\}_{n=1}^N$ samples from (x, y).

KCCA: notes

- Optimization domain: $\mathcal{H}_k \times \mathcal{H}_\ell \ni (f, g)$.
- By reproducing property: we will get a finite-D task.
- k,ℓ linear: traditional CCA.
- In practice: we have $\{(x_n, y_n)\}_{n=1}^N$ samples from (x, y).

Recall the reproducing property

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} \quad \forall f \in \mathcal{H}_k, x \in \mathcal{X}.$$

KCCA: empirical estimate

$$\widehat{\text{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\sum_{n=1}^{N}\Bigg[\underbrace{f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}}\Bigg]\Bigg[\underbrace{g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}\Bigg]$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\sum_{n=1}^{N}\Big[\underbrace{f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}}\Big]\Big[\underbrace{g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}\Big]$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

KCCA: empirical estimate

$$\widehat{\text{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\sum_{n=1}^{N}\Bigg[\underbrace{f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}}\Bigg]\Bigg[\underbrace{g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}\Bigg]$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\text{var}}_x f(x) = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}_{xy}(f(x),g(y)) = \frac{1}{N}\sum_{n=1}^{N}\Bigg[\underbrace{f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)}_{\left\langle f,\varphi(x_n)-\frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}}\Bigg]\Bigg[\underbrace{g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)}_{\left\langle g,\psi(y_n)-\frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}\Bigg]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\langle f,\tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}\langle g,\tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\mathrm{var}}_x f(x) = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2 = \frac{1}{N}\sum_{n=1}^{N}\langle f,\tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}^2,$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\sum_{n=1}^{N}\Bigg[\underbrace{f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}}\Bigg]\Bigg[\underbrace{g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}\Bigg]$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\mathrm{var}}_x f(x) = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2 = \frac{1}{N}\sum_{n=1}^{N}\langle f, \tilde{\varphi}(x_n)\rangle^2_{\mathcal{H}_k},$$

$$\widehat{\mathrm{var}}_y g(y) = \frac{1}{N}\sum_{n=1}^{N}\langle g, \tilde{\psi}(y_n)\rangle^2_{\mathcal{H}_\ell}.$$

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ component of f $\perp$

$$span\left(\{\tilde{\varphi}(x_n)\}_{n=1}^{N}\right) = \left\{\sum_{n=1}^{N} c_n \tilde{\varphi}(x_n), \mathbf{c} = [c_n] \in \mathbb{R}^N\right\}$$

has no affect in the objective.

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ component of f $\perp$

$$span\left(\{\tilde{\varphi}(x_n)\}_{n=1}^N\right) = \left\{\sum_{n=1}^N c_n \tilde{\varphi}(x_n), \mathbf{c} = [c_n] \in \mathbb{R}^N\right\}$$

has no affect in the objective.

Key idea

Enough to consider $f = \sum_{i=1}^N c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i \tilde{\psi}(y_i)$.

KCCA: empirical estimate

Using that $f = \sum_{i=1}^N c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k}$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \tilde{k}(x_i, x_n)$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \tilde{k}(x_i, x_n) = (\mathbf{c}^T \tilde{\mathbf{G}}_x)_n,$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^N c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \tilde{k}(x_i, x_n) = (\mathbf{c}^T \tilde{\mathbf{G}}_x)_n,$$

$$\left\langle g, \tilde{\psi}(y_n) \right\rangle_{\mathcal{H}_\ell} = (\mathbf{d}^T \tilde{\mathbf{G}}_y)_n,$$

with the centered kernels ($\tilde{k}$, $\tilde{\ell}$) and Gram matrices ($\tilde{\mathbf{G}}_x$, $\tilde{\mathbf{G}}_y$).

Until now

All the objective terms can be expressed by $\mathbf{c}$, $\mathbf{d}$, $\tilde{\mathbf{G}}_x$, $\tilde{\mathbf{G}}_y$.

KCCA: empirical estimate

$$\widehat{\text{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell},$$

$$\widehat{\text{var}}_x f(x) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}{}^2, \widehat{\text{var}}_y g(y) = \frac{1}{N}\sum_{n=1}^{N} \left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell}{}^2,$$

and we have

$$\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} = (\mathbf{c}^T \tilde{\mathbf{G}}_x)_n, \qquad \left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell} = (\mathbf{d}^T \tilde{\mathbf{G}}_y)_n.$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell},$$

$$\widehat{\mathrm{var}}_x f(x) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}{}^2, \widehat{\mathrm{var}}_y g(y) = \frac{1}{N}\sum_{n=1}^{N} \left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell}{}^2,$$

and we have

$$\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} = (\mathbf{c}^T\tilde{\mathbf{G}}_x)_n, \qquad \left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell} = (\mathbf{d}^T\tilde{\mathbf{G}}_y)_n.$$

Thus,

$$\widehat{\mathrm{cov}}_{xy}(f(x), g(y)) = \frac{1}{N}\mathbf{c}^T\tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y\mathbf{d},$$

$$\widehat{\mathrm{var}}_x f(x) = \frac{1}{N}\mathbf{c}^T(\tilde{\mathbf{G}}_x)^2\mathbf{c}, \quad \widehat{\mathrm{var}}_y g(y) = \frac{1}{N}\mathbf{d}^T(\tilde{\mathbf{G}}_y)^2\mathbf{d}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

In practice ($\kappa > 0$):

$$\widehat{\rho_{\text{KCCA}}}(x, y) := \widehat{\rho_{\text{KCCA}}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa)$$

$$= \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T \left(\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N\right)^2 \mathbf{c}} \sqrt{\mathbf{d}^T \left(\tilde{\mathbf{G}}_y + \kappa \mathbf{I}_N\right)^2 \mathbf{d}}}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

In practice ($\kappa > 0$):

$$\widehat{\rho_{\text{KCCA}}}(x, y) := \widehat{\rho_{\text{KCCA}}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa)$$

$$= \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T \left(\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N\right)^2 \mathbf{c}} \sqrt{\mathbf{d}^T \left(\tilde{\mathbf{G}}_y + \kappa \mathbf{I}_N\right)^2 \mathbf{d}}}.$$

Question

How do we solve it?

KCCA: solution

Stationary points of $\widehat{\rho_{\text{KCCA}}}(x,y)$:

$$\mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(x,y)}{\partial \mathbf{c}}, \qquad \mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(x,y)}{\partial \mathbf{d}},$$

which simplifies to

$$\tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y\mathbf{d} = \frac{(\mathbf{c}^T\tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y\mathbf{d})(\tilde{\mathbf{G}}_x+\kappa\mathbf{I}_N)^2\mathbf{c}}{\mathbf{c}^T(\tilde{\mathbf{G}}_x+\kappa\mathbf{I}_N)^2\mathbf{c}}, \quad \tilde{\mathbf{G}}_y\tilde{\mathbf{G}}_x\mathbf{c} = \frac{(\mathbf{d}^T\tilde{\mathbf{G}}_y\tilde{\mathbf{G}}_x\mathbf{c})(\tilde{\mathbf{G}}_y+\kappa\mathbf{I}_N)^2\mathbf{d}}{\mathbf{d}^T(\tilde{\mathbf{G}}_y+\kappa\mathbf{I}_N)^2\mathbf{d}}.$$

KCCA: solution

Stationary points of $\widehat{\rho_{\text{KCCA}}}(x,y)$:

$$\mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(x,y)}{\partial \mathbf{c}}, \qquad \mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(x,y)}{\partial \mathbf{d}},$$

which simplifies to

$$\tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d} = \frac{(\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d})(\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N)^2 \mathbf{c}}{\mathbf{c}^T (\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N)^2 \mathbf{c}}, \quad \tilde{\mathbf{G}}_y \tilde{\mathbf{G}}_x \mathbf{c} = \frac{(\mathbf{d}^T \tilde{\mathbf{G}}_y \tilde{\mathbf{G}}_x \mathbf{c})(\tilde{\mathbf{G}}_y + \kappa \mathbf{I}_N)^2 \mathbf{d}}{\mathbf{d}^T (\tilde{\mathbf{G}}_y + \kappa \mathbf{I}_N)^2 \mathbf{d}}.$$

Normalization:

- $(\mathbf{c}, \mathbf{d})$: solution $\Rightarrow$ $(a\mathbf{c}, b\mathbf{d})$: solution $a, b \in \mathbb{R}, \neq 0$.
- denominators $:= 1$.

KCCA: final task

Find the maximal eigenvalue, $\lambda := \mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}$, of the generalized eigenvalue problem:

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \\ \tilde{\mathbf{G}}_y \tilde{\mathbf{G}}_x & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_y + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$$\mathbf{A}\mathbf{z} = \lambda \mathbf{B}\mathbf{z}.$$

KCCA as an independence measure

If $x \perp y$, then $\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005b].

KCCA as an independence measure

If $x \perp y$, then $\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005b].
- Enough: universal kernel on a compact metric domain (later).

KCCA as an independence measure

If $x \perp y$, then $\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$ [Bach and Jordan, 2002, Gretton et al., 2005b].
- Enough: universal kernel on a compact metric domain (later).
- Example ($\gamma > 0$):
  - Gaussian: $k(x, x') = e^{-\gamma \|x - x'\|_2^2}$.
  - Laplacian kernel: $k(x, x') = e^{-\gamma \|x - x'\|_2}$.

KCCA: regularization

In fact, we estimated

$$\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y); \kappa),$$

$$\operatorname{corr}(f(x), g(y); \kappa) = \frac{\operatorname{cov}_{xy}(f(x), g(y))}{\sqrt{\operatorname{var}_x f(x) + \kappa \|f\|^2_{\mathcal{H}_k}}\sqrt{\operatorname{var}_y g(y) + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

KCCA: regularization

In fact, we estimated

$$\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \mathrm{corr}(f(x), g(y); \kappa),$$

$$\mathrm{corr}(f(x), g(y); \kappa) = \frac{\mathrm{cov}_{xy}(f(x), g(y))}{\sqrt{\mathrm{var}_x f(x) + \kappa \|f\|^2_{\mathcal{H}_k}} \sqrt{\mathrm{var}_y g(y) + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

- Regularization is important: With $\kappa = 0$, $\lambda \in \{0, \pm 1\} \Rightarrow$

$$\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 1$$

  would be data-independently [Gretton et al., 2005b], [Bach and Jordan, 2002].

KCCA: regularization

In fact, we estimated

$$\rho_{\text{KCCA}}(x, y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = \sup_{f\in\mathcal{H}_k, g\in\mathcal{H}_\ell} \operatorname{corr}(f(x), g(y); \kappa),$$

$$\operatorname{corr}(f(x), g(y); \kappa) = \frac{\operatorname{cov}_{xy}(f(x), g(y))}{\sqrt{\operatorname{var}_x f(x) + \kappa \|f\|^2_{\mathcal{H}_k}}\sqrt{\operatorname{var}_y g(y) + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

- For consistent KCCA estimate:
  - $\kappa_N \to 0$ [Leurgans et al., 1993](spline-RKHS), [Fukumizu et al., 2007] (general RKHS).
  - analysis: covariance operators (later).

KCCA: symmetry, other form

For a

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y \\ \tilde{\mathbf{G}}_y\tilde{\mathbf{G}}_x & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T\tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y\mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_x + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_y + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}.$$

$([\mathbf{c}, \mathbf{d}], \lambda)$ solution $\Rightarrow$ $([-\mathbf{c}; \mathbf{d}], -\lambda)$: solution. Thus, eigenvalues:

$$\{\lambda_1, -\lambda_1, \ldots, \lambda_N, -\lambda_N\}.$$

KCCA: symmetry, other form

For a

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y \\ \tilde{\mathbf{G}}_y\tilde{\mathbf{G}}_x & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T\tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y\mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_x + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_y + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}.$$

$([\mathbf{c}, \mathbf{d}], \lambda)$ solution $\Rightarrow$ $([-\mathbf{c}; \mathbf{d}], -\lambda)$: solution. Thus, eigenvalues:

$$\{\lambda_1, -\lambda_1, \ldots, \lambda_N, -\lambda_N\}.$$

Adding the r.h.s. to both sides:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_x + \kappa\mathbf{I}_N)^2 & \tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y \\ \tilde{\mathbf{G}}_y\tilde{\mathbf{G}}_x & (\tilde{\mathbf{G}}_\mathbf{y} + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_x + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_x + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

with eigenvalues $\{1+\lambda_1, 1-\lambda_1, \ldots, 1+\lambda_N, 1-\lambda_N\}$.

KCCA: M-variables

2-variables $[(x,y)]$:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_{\mathbf{x}}+\kappa\mathbf{I}_N)^2 & \tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y \\ \tilde{\mathbf{G}}_y\tilde{\mathbf{G}}_x & (\tilde{\mathbf{G}}_{\mathbf{y}}+\kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_x+\kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_x+\kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

KCCA: M-variables

2-variables $[(x,y)]$:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_{\mathbf{x}} + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \\ \tilde{\mathbf{G}}_y \tilde{\mathbf{G}}_x & (\tilde{\mathbf{G}}_{\mathbf{y}} + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_x + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

For M-variables (pairwise dependence):

$$\begin{bmatrix} (\tilde{\mathbf{G}}_1 + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_1 \tilde{\mathbf{G}}_2 & \dots & \tilde{\mathbf{G}}_1 \tilde{\mathbf{G}}_M \\ \tilde{\mathbf{G}}_2 \tilde{\mathbf{G}}_1 & (\tilde{\mathbf{G}}_2 + \kappa \mathbf{I}_N)^2 & \dots & \tilde{\mathbf{G}}_2 \tilde{\mathbf{G}}_M \\ \vdots & \vdots & & \vdots \\ \tilde{\mathbf{G}}_M \tilde{\mathbf{G}}_1 & \tilde{\mathbf{G}}_M \tilde{\mathbf{G}}_2 & \dots & (\tilde{\mathbf{G}}_M + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_M \end{bmatrix} =$$

$$\gamma \begin{bmatrix} (\tilde{\mathbf{G}}_1 + \kappa \mathbf{I}_N)^2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_2 + \kappa \mathbf{I}_N)^2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & \dots & (\tilde{\mathbf{G}}_M + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_M \end{bmatrix}.$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$(\tilde{\mathbf{G}}_x)_{ij} = \tilde{k}(x_i, x_j) = \left\langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}_k}$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N \times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_x)_{ij} &= \tilde{k}(x_i, x_j) = \left\langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k}
\end{aligned}
$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_x)_{ij} &= \tilde{k}(x_i, x_j) = \left\langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_x)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_x)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_x)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_x)_{nm}
\end{aligned}
$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N \times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_x)_{ij} &= \tilde{k}(x_i, x_j) = \left\langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_x)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_x)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_x)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_x)_{nm} \\
&= \left(\mathbf{G}_x - \mathbf{G}_x \frac{\mathbf{E}_N}{N} - \frac{\mathbf{E}_N}{N}\mathbf{G}_x - \frac{\mathbf{E}_N}{N}\mathbf{G}_x\frac{\mathbf{E}_N}{N} \right)_{ij},
\end{aligned}
$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_x)_{ij} &= \tilde{k}(x_i, x_j) = \left\langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_x)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_x)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_x)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_x)_{nm} \\
&= \left(\mathbf{G}_x - \mathbf{G}_x \frac{\mathbf{E}_N}{N} - \frac{\mathbf{E}_N}{N}\mathbf{G}_x - \frac{\mathbf{E}_N}{N}\mathbf{G}_x\frac{\mathbf{E}_N}{N} \right)_{ij}, \\
&= (\mathbf{H}\mathbf{G}_x\mathbf{H})_{ij},
\end{aligned}
$$

$\mathbf{H}$: symmetric ($\mathbf{H} = \mathbf{H}^T$), idempotent ($\mathbf{H}^2 = \mathbf{H}$).

KCCA: finished.

Mean embedding

Mean embedding: pioneers

- Nonparametric probability distribution representation.
- Late 70s-; survey in [Berlinet and Thomas-Agnan, 2004].

Mean embedding: pioneers

- Nonparametric probability distribution representation.
- Late 70s-; survey in [Berlinet and Thomas-Agnan, 2004].
- Pioneers in ML: Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Alex Smola, Bernhard Schölkopf, Le Song.

Mean embedding: further pointers

- Names+: Ingo Steinwart, Francis Bach, Dino Sejdinovic, Wittawat Jitkrittum, Krikamol Maundet, Kacper P. Chwialkowski, Ilya Tolstikhin, Carl Johann Simon-Gabriel, David Lopez-Paz, Dougal Sutherland, Aaditya Ramdas, Karsten Borgwardt, Me;)

Mean embedding: further pointers

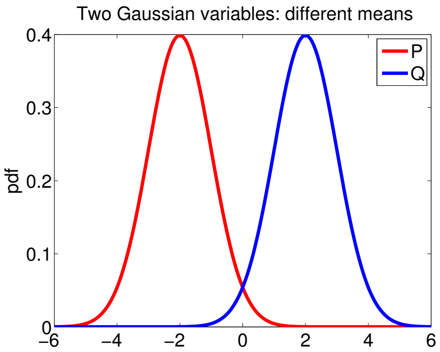
- Names+: Ingo Steinwart, Francis Bach, Dino Sejdinovic, Wittawat Jitkrittum, Krikamol Maundet, Kacper P. Chwialkowski, Ilya Tolstikhin, Carl Johann Simon-Gabriel, David Lopez-Paz, Dougal Sutherland, Aaditya Ramdas, Karsten Borgwardt, Me;)
- Wiki: `https://en.wikipedia.org/wiki/Kernel_embedding_of_distributions`.

Mean embedding: further pointers

- Names+: Ingo Steinwart, Francis Bach, Dino Sejdinovic, Wittawat Jitkrittum, Krikamol Maundet, Kacper P. Chwialkowski, Ilya Tolstikhin, Carl Johann Simon-Gabriel, David Lopez-Paz, Dougal Sutherland, Aaditya Ramdas, Karsten Borgwardt, Me;)
- Wiki: `https://en.wikipedia.org/wiki/Kernel_embedding_of_distributions`.
- Recent review: [Muandet et al., 2017].

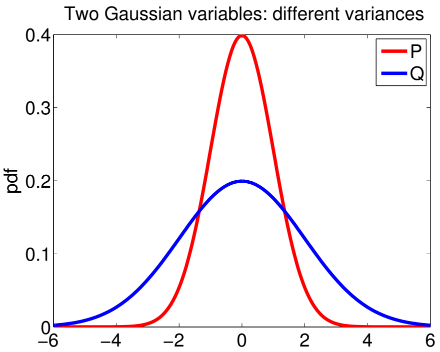
Towards representations of distributions: $\mathbb{E}X$

- Given: 2 Gaussians with different means.
- Solution: *t*-test.



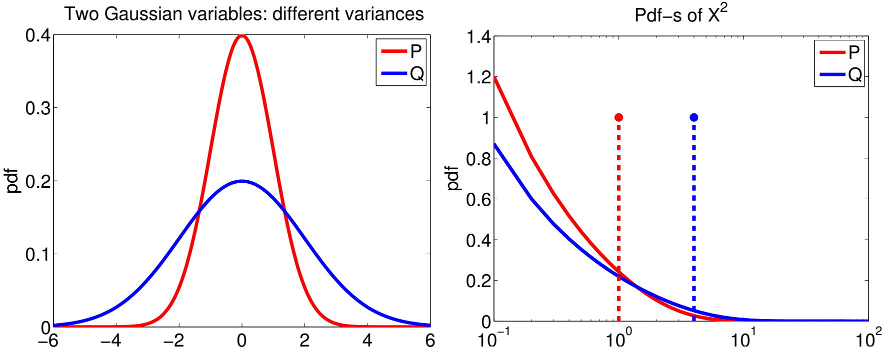
Towards representations of distributions: $\mathbb{E}X^2$

- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.



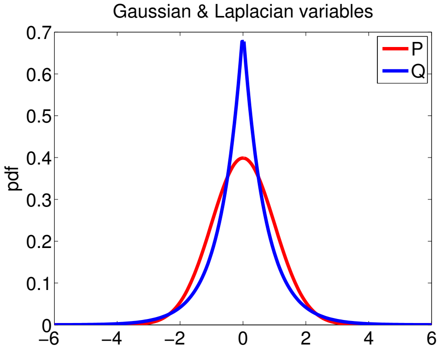
Towards representations of distributions: $\mathbb{E}X^2$

- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.
- $\varphi_x = x^2 \Rightarrow$ difference in $\mathbb{E}X^2$.



Towards representations of distributions: further moments

- Setup: a Gaussian and a Laplacian distribution.
- Challenge: their means *and* variances are the same.
- Idea: look at higher-order features.



Let us consider feature representations!

From kernel trick to mean trick

- Recall:
  - $\varphi(x) \in \mathcal{H}_k$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}_k}$.

From kernel trick to mean trick

- Recall:
  - $\varphi(x) \in \mathcal{H}_k$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}_k}$.
- Mean embedding:
  - Feature of $\mathbb{P}$:

$$\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}}[\varphi(x)] \in \mathcal{H}_k.$$

  - Inner product: $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \mathbb{E}_{x \sim \mathbb{P}, x' \sim \mathbb{Q}} k(x, x')$.

From kernel trick to mean trick

- Recall:
  - $\varphi(x) \in \mathcal{H}_k$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}_k}$.
- Mean embedding:
  - Feature of $\mathbb{P}$:

$$\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}}[\varphi(x)] \in \mathcal{H}_k.$$

  - Inner product: $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \mathbb{E}_{x \sim \mathbb{P}, x' \sim \mathbb{Q}} k(x, x')$.
- $\mu_{\mathbb{P}}$: well-defined for all distributions (bounded k).

From kernel trick to mean trick

- Recall:
  - $\varphi(x) \in \mathcal{H}_k$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}_k}$.
- Mean embedding:
  - Feature of $\mathbb{P}$:

$$\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}}[\varphi(x)] \in \mathcal{H}_k.$$

  - Inner product: $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \mathbb{E}_{x \sim \mathbb{P}, x' \sim \mathbb{Q}} k(x, x')$.
- $\mu_{\mathbb{P}}$: well-defined for all distributions (bounded k).

Commonly used construction

$\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}}[\varphi(x)]$. Indeed. . .

Distribution Representation via Functions

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z)$$

Distribution Representation via Functions

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

Distribution Representation via Functions

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto \phi_{\mathbb{P}}(z) = \int e^{i\langle z, x\rangle} \mathrm{d}\mathbb{P}(x).$$

Distribution Representation via Functions

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}}\chi_{(-\infty,z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto \phi_{\mathbb{P}}(z) = \int e^{i\langle z,x\rangle}\mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z,x\rangle}\mathrm{d}\mathbb{P}\left(x\right).$$

Distribution Representation via Functions

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}}\chi_{(-\infty,z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto \phi_{\mathbb{P}}(z) = \int e^{i\langle z,x\rangle}\mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z,x\rangle}\mathrm{d}\mathbb{P}\left(x\right).$$

Pattern

$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x)\mathrm{d}\mathbb{P}(x)$

Distribution Representation via Functions

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto \phi_{\mathbb{P}}(z) = \int e^{i\langle z, x \rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z, x \rangle} \mathrm{d}\mathbb{P}(x).$$

Pattern

$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x)$, in our case: $\varphi(x) = k(\cdot, x)$.

Bochner integral: quick summary [Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to B$(anach space)-valued measurable function.

Bochner integral: quick summary [Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to B$(anach space)-valued measurable function.
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ ($A_i \in \mathcal{A}, c_i \in B$) measurable step functions

$$\int_{\mathcal{X}} f \mathrm{d}\mu := \sum_{i=1}^{n} c_i \mu(A_i) \in B.$$

Bochner integral: quick summary [Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to B$(anach space)-valued measurable function.
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in B)$ measurable step functions

$$\int_{\mathcal{X}} f \mathrm{d}\mu := \sum_{i=1}^{n} c_i \mu(A_i) \in B.$$

- f measurable function is Bochner μ-integrable if
  - $\exists$ (f_n) measurable step functions: $\lim_{n\to\infty} \int_{\mathcal{X}} \|f - f_n\|_B \mathrm{d}\mu = 0$.
  - In this case $\lim_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\mu$ exists, $=: \int_{\mathcal{X}} f \mathrm{d}\mu$.

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \,\mathrm{d}\mu < \infty$.

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \mathrm{d}\mu\right\|_B \leqslant \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu$. ('Jensen inequality')

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \mathrm{d}\mu\right\|_B \leqslant \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu$. ('Jensen inequality')
- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

  $S \circ f : X \to B_2$ is Bochner integrable and

$$S\left(\int_{\mathcal{X}} f \mathrm{d}\mu\right) = \int_{\mathcal{X}} S f \mathrm{d}\mu.$$

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \mathrm{d}\mu\right\|_B \leqslant \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu$. ('Jensen inequality')
- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

  $S \circ f : X \to B_2$ is Bochner integrable and

$$S\left(\int_{\mathcal{X}} f \mathrm{d}\mu\right) = \int_{\mathcal{X}} S f \mathrm{d}\mu.$$

In short

$\left|\int f \mathrm{d}\mu\right| \leqslant \int |f| \mathrm{d}\mu$ and $c \int f \mathrm{d}\mu = \int c f \mathrm{d}\mu$ generalize nicely.

Mean embedding: $\exists$, $\mathbb{E}_{\mathbb{P}}$-reproducing property

Given:

- $(\mathcal{X}, \mathcal{A})$ measurable space,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel.

Theorem

$\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x)$ *exists*, $\mu_{\mathbb{P}} \in \mathcal{H}_k$, *and*

$$\mathbb{P}f := \mathbb{E}_{x\sim\mathbb{P}} f(x) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_k} \ \forall f \in \mathcal{H}_k$$

under mild conditions:

- $\mathbb{E}_{x\sim\mathbb{P}} \sqrt{k(x,x)} < \infty$, *and*
- $y \mapsto k(y, x)$ *is measurable for any* $x \in \mathcal{X}$.

Existence of $\mu_{\mathbb{P}}$: proof

- $\exists \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x)$ ($\& \in \mathcal{H}_k$) $\Leftrightarrow$

$$\infty > \int_{\mathcal{X}} \|k(\cdot,x)\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) = \mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)}.$$

Existence of $\mu_{\mathbb{P}}$: proof

- $\exists \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x)$ ($\& \in \mathcal{H}_k$) $\Leftrightarrow$

$$\infty > \int_{\mathcal{X}} \|k(\cdot,x)\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) = \mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)}.$$

- $\mathbb{E}_{x\sim\mathbb{P}} f(x) = \mathbb{E}_{x\sim\mathbb{P}} \langle f, k(\cdot,x)\rangle_{\mathcal{H}_k} = \langle f, \mathbb{E}_{x\sim\mathbb{P}} k(\cdot,x)\rangle_{\mathcal{H}_k} = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k}$ by
  - reproducing property of k,
  - $g \in \mathcal{H}_k \mapsto \langle f, g\rangle \in \mathbb{R}$: bounded linear ($S \leftrightarrow \int$).

Existence of $\mu_{\mathbb{P}}$: proof

- $\exists \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x)$ ($\& \in \mathcal{H}_k$) $\Leftrightarrow$

$$\infty > \int_{\mathcal{X}} \|k(\cdot,x)\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) = \mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)}.$$

- $\mathbb{E}_{x\sim\mathbb{P}} f(x) = \mathbb{E}_{x\sim\mathbb{P}} \langle f, k(\cdot,x)\rangle_{\mathcal{H}_k} = \langle f, \mathbb{E}_{x\sim\mathbb{P}} k(\cdot,x)\rangle_{\mathcal{H}_k} = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k}$ by
  - reproducing property of k,
  - $g \in \mathcal{H}_k \mapsto \langle f, g\rangle \in \mathbb{R}$: bounded linear ($S \leftrightarrow \int$).
- Measurability of $x \in \mathcal{X} \mapsto k(\cdot,x) \in \mathcal{H}_k$: $\Leftrightarrow$ $y \mapsto k(y,x)$ is measurable $\forall x$ [Berlinet and Thomas-Agnan, 2004].

Mean embedding: specific cases

For

- $k(x, x') = e^{\langle x, x' \rangle}$: $\mu_{\mathbb{P}}$ = moment generating function of $\mathbb{P}$.
- $k(x, y) = e^{i\langle x, y \rangle}$: $\mu_{\mathbb{P}}$ = characteristic function of $\mathbb{P}$.
  - Only formally: $k(x, y) = k(y, x)^*$ fails.
- $\mathbb{P} = \delta_x$, $\mu_{\mathbb{P}} = k(\cdot, x)$.

Mean embedding: conditions

Condition:

- $y \mapsto k(y,x)$ is measurable $\forall x$: super-mild.
- $\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)} < \infty$: holds for bounded kernels, i.e. when

$$\sup_{x,x'\in\mathcal{X}} k(x,x') \leqslant B_k < \infty.$$

Mean embedding: empirical estimate

- $\mu_{\mathbb{P}}$: typically analytically not available.
- Empirical estimate: from $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$

$$\widehat{\mu_{\mathbb{P}}} = \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i) = \mu_{\mathbb{P}_n} \in \mathcal{H}_k,$$

where $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical measure.

Empirical mean embedding: finite-sample guarantees

Theorem ([Altun and Smola, 2006])

For a k bounded kernel [$\sup_{x,y\in\mathcal{X}} k(x,y) \leqslant B_k$], with probability $\geqslant 1-\delta$

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{P}_n}\|_{\mathcal{H}_k} \leqslant \frac{\left[1+\sqrt{\log\left(\frac{1}{\delta}\right)}\right]\sqrt{2B_k}}{\sqrt{n}}.$$

Finite-sample guarantee: proof idea

- $g(x_1, \ldots, x_n) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{P}_n}\|_{\mathcal{H}_k}$: bounded difference property $\Rightarrow$
- McDiarmid inequality: concentration around $\mathbb{E}g$.
- $\mathbb{E}g \leqslant$ expected kernel values (B_k appears).

Finite-sample guarantee: note

Alternative of

$$\mathbb{P}\left(\|\mu_{\mathbb{P}} - \mu_{\mathbb{P}_n}\|_{\mathcal{H}_k} \leqslant \frac{\left[1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right]\sqrt{2B_k}}{\sqrt{n}}\right) \geqslant 1 - \delta.$$

Directly by the Bernstein inequality [Caponnetto and De Vito, 2007]:

$$\mathbb{P}\left(\|\mu_{\mathbb{P}} - \mu_{\mathbb{P}_n}\|_{\mathcal{H}_k} \leqslant 2\sqrt{B_k}\left[\frac{2}{n} + \frac{1}{\sqrt{n}}\log\left(\frac{2}{\delta}\right)\right]\right) \geqslant 1 - \delta$$

would give a bit worse dependence.

MMD: preview

- Mean embeddings define a semi-metric (MMD):

$$d_k(\mathbb{P},\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

MMD: preview

- Mean embeddings define a semi-metric (MMD):

$$d_k(\mathbb{P}, \mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

- d_k is metric $\Leftrightarrow \mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective.
- Characteristic kernel [Fukumizu et al., 2004, Fukumizu et al., 2008]:
  - characteristic function analogy.
  - L-order polynomial kernel: encodes moments $\leqslant L$. (not)

Mean embedding: universality (k)

Universal kernel

Let $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$.

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Universal kernel

Let $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$.

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

$\mathcal{X}$ assumption $\Rightarrow$

$C(\mathcal{X}) = C_b(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous bounded}\}$

Universal kernel

$\mathcal{H}_k \subset C(\mathcal{X})$? Non-compact spaces?

Notes:

- k: continuous, $\mathcal{X}$: compact $\Rightarrow$ k: bounded.
- k: continuous, bounded $\Rightarrow$ $\mathcal{H}_k \subset C(\mathcal{X})$ [Steinwart and Christmann, 2008].

Universal kernel

$\mathcal{H}_k \subset C(\mathcal{X})$? Non-compact spaces?

Notes:

- Extensions of c-universality to non-compact spaces:
  - c_0-universality, cc-universality, ... [Carmeli et al., 2010, Sriperumbudur et al., 2010a, Simon-Gabriel and Schölkopf, 2016].

k: universal $\Rightarrow$ k: characteristic

$\geqslant 3$ different proof options:

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.

k: universal $\Rightarrow$ k: characteristic

$\geqslant 3$ different proof options:

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.
- Direct reasoning [Gretton et al., 2012].

k: universal $\Rightarrow$ k: characteristic

$\geqslant 3$ different proof options:

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.
- Direct reasoning [Gretton et al., 2012].
- Denseness of $\mathcal{H}_k + \mathbb{R}$ in $L^2(\mathbb{P})$ [Fukumizu et al., 2008, Fukumizu et al., 2009a].

k: universal $\Rightarrow$ k: characteristic

$\geqslant 3$ different proof options:

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.
- Direct reasoning [Gretton et al., 2012].
- Denseness of $\mathcal{H}_k + \mathbb{R}$ in $L^2(\mathbb{P})$ [Fukumizu et al., 2008, Fukumizu et al., 2009a].

Let us construct some *examples* first! (then proof 1-2)

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\varphi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x,y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}_k}$$

is a metric.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\varphi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x,y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}_k}$$

  is a metric.
- The normalized kernel

$$\tilde{k}(x,y) := \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

  is universal.

Universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^{\infty} \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r), r \in (0, \infty].$$

Universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^{\infty} \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r), r \in (0, \infty].$$

- If $a_n > 0\ \forall n$, then

$$k(x, y) = f(\langle x, y \rangle)$$

is universal on $\mathcal{X} := \left\{ x \in \mathbb{R}^d : \|x\|_2 \leqslant \sqrt{r} \right\}$.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(x, y) = e^{\alpha\langle x,y\rangle}$: previous result with $a_n = \frac{\alpha^n}{n!}$.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(x, y) = e^{\alpha\langle x,y\rangle}$: previous result with $a_n = \frac{\alpha^n}{n!}$.
- $k(x, y) = e^{-\alpha\|x-y\|_2^2}$: exp. kernel & normalization.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(x,y) = (1 - \langle x, y\rangle)^{-\alpha}$ binomial kernel
  - on $\mathcal{X}$ compact $\subset \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$.
  - $f(t) = (1-t)^{-\alpha} = \sum_{n=0}^{\infty} \underbrace{\binom{-\alpha}{n}(-1)^n}_{>0} t^n \;(|t| < 1)$,

where $\binom{b}{n} = \sum_{i=1}^{n} \frac{b-i+1}{i}$.

Universal $\Rightarrow$ characteristic: proof-1

Injectivity on finite signed measures (proof):

- k: universal $\Rightarrow$ $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.

Universal $\Rightarrow$ characteristic: proof-1

Injectivity on finite signed measures (proof):

- k: universal $\Rightarrow$ $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.
- Hahn-Banach theorem [Rudin, 1991]: Let H is a subspace of a normed space C. H is dense in C iff.

$$\{0\} = H^{\perp} := \{F \in C' : \forall f \in H, F(f) = 0\}.$$

Universal $\Rightarrow$ characteristic: proof-1

Injectivity on finite signed measures (proof):

- k: universal $\Rightarrow$ $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.
- Hahn-Banach theorem [Rudin, 1991]: Let H is a subspace of a normed space C. H is dense in C iff.

$$\{0\} = H^{\perp} := \{F \in C' : \forall f \in H, F(f) = 0\}.$$

- Denseness $\Leftrightarrow$

$$\{0\} = \mathcal{H}_k^{\perp} = \left\{\mathbb{F} \in \underbrace{C(\mathcal{X})'}_{=\mathcal{M}_b(\mathcal{X})} : \forall f \in \mathcal{H}_k, 0 = T_{\mathbb{F}}(f) = \underbrace{\int_{\mathcal{X}} f \mathrm{d}\mathbb{F}}_{\langle f, \mu_{\mathbb{F}} \rangle_{\mathcal{H}_k}}\right\}$$

Universal $\Rightarrow$ characteristic: proof-1

Injectivity on finite signed measures (proof):

- k: universal $\Rightarrow \mathcal{H}_k$ is dense in $C(\mathcal{X})$.
- Hahn-Banach theorem [Rudin, 1991]: Let H is a subspace of a normed space C. H is dense in C iff.

$$\{0\} = H^{\perp} := \{F \in C' : \forall f \in H, F(f) = 0\}.$$

- Denseness $\Leftrightarrow$

$$\{0\} = \mathcal{H}_k^{\perp} = \left\{\mathbb{F} \in \underbrace{C(\mathcal{X})'}_{=\mathcal{M}_b(\mathcal{X})} : \forall f \in \mathcal{H}_k, 0 = T_{\mathbb{F}}(f) = \underbrace{\int_{\mathcal{X}} f \mathrm{d}\mathbb{F}}_{\langle f, \mu_{\mathbb{F}} \rangle_{\mathcal{H}_k}}\right\}$$

$$= \{\mathbb{F} \in \mathcal{M}_b(\mathcal{X}) : \mu_{\mathbb{F}} = 0\}.$$

Universal ⇒ characteristic: proof-2

Direct reasoning: We have already mentioned [Dudley, 2004]:

- Let $\mathcal{X}$: metric space, $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$.
- Then $\mathbb{P} = \mathbb{Q}$ (Borel probability measures) ⇔

$$\mathbb{P}f = \mathbb{Q}f \quad \forall f \in C_b(\mathcal{X}).$$

We have a characterization of $\mathbb{P} = \mathbb{Q}$ in terms of expectations.

Universal $\Rightarrow$ characteristic: proof-2

- Goal: $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$ [$\Leftrightarrow \mathbb{P}f = \mathbb{Q}f, \forall f \in C_b(\mathcal{X})$].

Universal $\Rightarrow$ characteristic: proof-2

- Goal: $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$ $[\Leftrightarrow \mathbb{P}f = \mathbb{Q}f, \forall f \in C_b(\mathcal{X})]$.
- We want: for any $f \in C_b(\mathcal{X})$ and $\epsilon > 0$, $|\mathbb{P}f - \mathbb{Q}f| \overset{?}{\leqslant} \epsilon$.

Universal $\Rightarrow$ characteristic: proof-2

- Goal: $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$ $[\Leftrightarrow \mathbb{P}f = \mathbb{Q}f, \forall f \in C_b(\mathcal{X})]$.
- We want: for any $f \in C_b(\mathcal{X})$ and $\epsilon > 0$, $|\mathbb{P}f - \mathbb{Q}f| \overset{?}{\leqslant} \epsilon$.
- Universality of $k \Rightarrow \mathcal{H}_k$ is dense in $C_b(\mathcal{X})$.

Universal $\Rightarrow$ characteristic: proof-2

- Goal: $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$ $[\Leftrightarrow \mathbb{P}f = \mathbb{Q}f, \forall f \in C_b(\mathcal{X})]$.
- We want: for any $f \in C_b(\mathcal{X})$ and $\epsilon > 0$, $|\mathbb{P}f - \mathbb{Q}f| \overset{?}{\leqslant} \epsilon$.
- Universality of $k \Rightarrow \mathcal{H}_k$ is dense in $C_b(\mathcal{X})$.
- $\mathcal{H}_k \ni$ g:= ϵ-approximation of f,

$$|\mathbb{P}f - \mathbb{Q}f| \leqslant \underbrace{|\mathbb{P}f - \mathbb{P}g|}_{\leqslant \mathbb{P}|f-g| \leqslant \epsilon} + |\mathbb{P}g - \mathbb{Q}g| + \underbrace{|\mathbb{Q}g - \mathbb{Q}f|}_{\leqslant \epsilon},$$

Universal $\Rightarrow$ characteristic: proof-2

- Goal: $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$ $[\Leftrightarrow \mathbb{P}f = \mathbb{Q}f, \forall f \in C_b(\mathcal{X})]$.
- We want: for any $f \in C_b(\mathcal{X})$ and $\epsilon > 0$, $|\mathbb{P}f - \mathbb{Q}f| \overset{?}{\leqslant} \epsilon$.
- Universality of $k \Rightarrow \mathcal{H}_k$ is dense in $C_b(\mathcal{X})$.
- $\mathcal{H}_k \ni$ g:= ϵ-approximation of f,

$$|\mathbb{P}f - \mathbb{Q}f| \leqslant \underbrace{|\mathbb{P}f - \mathbb{P}g|}_{\leqslant \mathbb{P}|f-g| \leqslant \epsilon} + |\mathbb{P}g - \mathbb{Q}g| + \underbrace{|\mathbb{Q}g - \mathbb{Q}f|}_{\leqslant \epsilon},$$

$$|\mathbb{P}g - \mathbb{Q}g| = |\underbrace{\langle g, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k} - \langle g, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}_k}}_{\left\langle g, \underbrace{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}_{=0} \right\rangle_{\mathcal{H}_k}}| = 0. \text{ Thus } |\mathbb{P}f - \mathbb{Q}f| \leqslant 2\epsilon.$$

Universality: finished. Now: characteristic property.

$d_k(\mathbb{P}, \mathbb{Q})$ (=MMD) in terms of kernel evaluations

[Gretton et al., 2007]:

$$d_k^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}^2 = \left\| \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, y) \mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2$$

$d_k(\mathbb{P},\mathbb{Q})$ (=MMD) in terms of kernel evaluations

[Gretton et al., 2007]:

$$d_k^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}^2 = \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2$$
$$= \langle a-b, a-b\rangle_{\mathcal{H}_k}$$

$d_k(\mathbb{P},\mathbb{Q})$ (=MMD) in terms of kernel evaluations

[Gretton et al., 2007]:

$$d_k^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k} = \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|^2_{\mathcal{H}_k}$$
$$= \langle a - b, a - b \rangle_{\mathcal{H}_k}$$
$$= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_k} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} - 2 \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k},$$

.

$d_k(\mathbb{P},\mathbb{Q})$ (=MMD) in terms of kernel evaluations

[Gretton et al., 2007]:

$$d_k^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}^2 = \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2$$

$$= \langle a-b, a-b\rangle_{\mathcal{H}_k}$$

$$= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}}\rangle_{\mathcal{H}_k} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}_k} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}_k},$$

$$= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{P}(x') + \int_{\mathcal{X}}\int_{\mathcal{X}} k(y,y')\mathrm{d}\mathbb{Q}(y)\mathrm{d}\mathbb{Q}(y')$$

$$- 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y)$$

.

$d_k(\mathbb{P}, \mathbb{Q})$ (=MMD) in terms of kernel evaluations

[Gretton et al., 2007]:

$$
\begin{aligned}
d_k^2(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}^2 = \left\| \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, y) \mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2 \\
&= \langle a - b, a - b \rangle_{\mathcal{H}_k} \\
&= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_k} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} - 2 \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}, \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') \mathrm{d}\mathbb{P}(x) \mathrm{d}\mathbb{P}(x') + \int_{\mathcal{X}} \int_{\mathcal{X}} k(y, y') \mathrm{d}\mathbb{Q}(y) \mathrm{d}\mathbb{Q}(y') \\
&\quad - 2 \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \mathrm{d}\mathbb{P}(x) \mathrm{d}\mathbb{Q}(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \mathrm{d}(\mathbb{P} - \mathbb{Q})(x) \mathrm{d}(\mathbb{P} - \mathbb{Q})(y).
\end{aligned}
$$

$\Rightarrow$ Polynomial kernels are *not* characteristic

[Sriperumbudur et al., 2010b]:

- $k(x,y) = \langle x, y \rangle$: linear kernel ($L = 1$).

$$d_k^2(\mathbb{P},\mathbb{Q}) = \left\| m_{\mathbb{P}} - m_{\mathbb{Q}} \right\|^2, \qquad m_{\mathbb{P}} = \int_{\mathcal{X}} x \mathrm{d}\mathbb{P}(x).$$

$\Rightarrow$ Polynomial kernels are *not* characteristic

[Sriperumbudur et al., 2010b]:

- $k(x,y) = \langle x, y\rangle$: linear kernel ($L = 1$).

$$d_k^2(\mathbb{P},\mathbb{Q}) = \|m_{\mathbb{P}} - m_{\mathbb{Q}}\|^2, \qquad m_{\mathbb{P}} = \int_{\mathcal{X}} x \mathrm{d}\mathbb{P}(x).$$

- $k(x,y) = (\langle x, y\rangle + 1)^2$ ($L = 2$):

$$d_k^2(\mathbb{P},\mathbb{Q}) = 2\|m_{\mathbb{P}} - m_{\mathbb{P}}\|^2 + \left\|\Sigma_{\mathbb{P}} - \Sigma_{\mathbb{Q}} + m_{\mathbb{P}} m_{\mathbb{P}}^T - m_{\mathbb{Q}} m_{\mathbb{Q}}^T\right\|_F^2,$$

where $\|\cdot\|_F$: Frobenious norm; $\Sigma_{\mathbb{P}}$: cov. matrix w.r.t. $\mathbb{P}$.

Characteristic property

Well-understood for

- Continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(x,y) = k_0(x-y), k_0 \in C_b(\mathbb{R}^d).$$

Characteristic property

Well-understood for

- Continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(x,y) = k_0(x-y), k_0 \in C_b(\mathbb{R}^d).$$

- Continuous bounded radial kernels on $\mathbb{R}^d$:

$$k(x,y) = k_0(\|x-y\|_2), \quad k_0 \in C_b(\mathbb{R}^d),$$
$$k_0(z) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2} \mathrm{d}\nu(t)$$

$\nu \in \mathcal{M}_b^+[0,\infty)$, i.e. it is a finite measure on $[0,\infty)$.

Bochner's theorem

We focus on continuous bounded shift-invariant kernels:

Theorem (Bochner's theorem [Wendland, 2005], $k \leftrightarrow \Lambda$)

$$k_0(z) = \int_{\mathbb{R}^d} e^{-i\langle z,\omega\rangle} \mathrm{d}\Lambda(\omega),$$

where Λ is a finite Borel measure (w.l.o.g. probability).

MMD in terms of characteristic functions

Using Bochner's theorem:

$$d_k^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(x, y) \mathrm{d}(\mathbb{P} - \mathbb{Q})(x) \mathrm{d}(\mathbb{P} - \mathbb{Q})(y)$$

MMD in terms of characteristic functions

Using Bochner's theorem:

$$\begin{aligned} d_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\mathrm{d}\Lambda(\omega)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \end{aligned}$$

MMD in terms of characteristic functions

Using Bochner's theorem:

$$\begin{aligned}
d_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\mathrm{d}\Lambda(\omega)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\\
&= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle x,\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\right]}_{\overline{\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle y,\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\right]}_{\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)}\mathrm{d}\Lambda(\omega)
\end{aligned}$$

MMD in terms of characteristic functions

Using Bochner's theorem:

$$
\begin{aligned}
d_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\mathrm{d}\Lambda(\omega)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle x,\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\right]}_{\overline{\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle y,\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\right]}_{\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)} \mathrm{d}\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \left|\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)\right|^2 \mathrm{d}\Lambda(\omega)
\end{aligned}
$$

MMD in terms of characteristic functions

Using Bochner's theorem:

$$
\begin{aligned}
d_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\mathrm{d}\Lambda(\omega)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\\
&= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle x,\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\right]}_{\overline{\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle y,\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\right]}_{\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)} \mathrm{d}\Lambda(\omega)\\
&= \int_{\mathbb{R}^d} \left|\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)\right|^2 \mathrm{d}\Lambda(\omega) = \left\|\phi_{\mathbb{P}}-\phi_{\mathbb{Q}}\right\|^2_{L^2(\Lambda)}.
\end{aligned}
$$

Continuous bounded translation-invariant kernels

Theorem ([Sriperumbudur et al., 2010b])

They are characteristic iff. $supp(\Lambda) = \mathbb{R}^d$.

Continuous bounded translation-invariant kernels

Theorem ([Sriperumbudur et al., 2010b])

They are characteristic iff. $supp(\Lambda) = \mathbb{R}^d$.

$supp(\Lambda) := \{x \in \mathcal{X}: \text{ for any open set } U \text{ such that } x \in U, \Lambda(U) > 0\}$.

Continuous bounded translation-invariant kernels

Theorem ([Sriperumbudur et al., 2010b])

They are characteristic iff. $supp(\Lambda) = \mathbb{R}^d$.

$supp(\Lambda) := \{x \in \mathcal{X}: \text{ for any open set } U \text{ such that } x \in U, \Lambda(U) > 0\}$.

- Example: Gaussian, Laplacian, Matérn kernel, B-spline kernel.

Continuous bounded translation-invariant kernels

Theorem ([Sriperumbudur et al., 2010b])

They are characteristic iff. $supp(\Lambda) = \mathbb{R}^d$.

$supp(\Lambda) := \{x \in \mathcal{X} \colon \text{ for any open set } U \text{ such that } x \in U, \Lambda(U) > 0\}$.

- Example: Gaussian, Laplacian, Matérn kernel, B-spline kernel.
- Similar characterization $\exists$ on 'Bochner domains' (LCA groups [Berg et al., 1984], orthogonal matrices, $\mathbb{R}^d_+$) [Fukumizu et al., 2009b].

Matérn kernel

$$k(x,y) = k_0(x-y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\left\|x-y\right\|_2}{\sigma}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}\left\|x-y\right\|_2}{\sigma}\right),$$

where K_ν: modified Bessel function of the second kind of order ν

Matérn kernel

$$k(x,y) = k_0(x-y) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,\|x-y\|_2}{\sigma}\right)^{v} K_v\left(\frac{\sqrt{2v}\,\|x-y\|_2}{\sigma}\right),$$

$$\widehat{k_0}(\omega) = \frac{2^{d+v}\pi^{\frac{d}{2}}\Gamma(v+d/2)v^v}{\Gamma(v)\sigma^{2v}}\left(\frac{2v}{\sigma^2} + 4\pi^2\|\omega\|_2^2\right)^{-(v+d/2)} > 0 \quad \forall \omega \in \mathbb{R}^d,$$

where K_v: modified Bessel function of the second kind of order v, Γ: Gamma function.

Matérn kernel

$$k(x,y) = k_0(x-y) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,\|x-y\|_2}{\sigma}\right)^{v} K_v\left(\frac{\sqrt{2v}\,\|x-y\|_2}{\sigma}\right),$$

$$\widehat{k_0}(\omega) = \frac{2^{d+v}\pi^{\frac{d}{2}}\Gamma(v+d/2)v^v}{\Gamma(v)\sigma^{2v}}\left(\frac{2v}{\sigma^2} + 4\pi^2\|\omega\|_2^2\right)^{-(v+d/2)} > 0 \quad \forall \omega \in \mathbb{R}^d,$$

where K_v: modified Bessel function of the second kind of order v, Γ: Gamma function.

- For $v = \frac{1}{2}$: one gets $k(x,y) = e^{-\frac{\|x-y\|_2}{\sigma}}$.

Matérn kernel

$$k(x,y) = k_0(x-y) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,\|x-y\|_2}{\sigma}\right)^{v} K_v\left(\frac{\sqrt{2v}\,\|x-y\|_2}{\sigma}\right),$$

$$\widehat{k_0}(\omega) = \frac{2^{d+v}\pi^{\frac{d}{2}}\Gamma(v+d/2)v^v}{\Gamma(v)\sigma^{2v}}\left(\frac{2v}{\sigma^2} + 4\pi^2\|\omega\|_2^2\right)^{-(v+d/2)} > 0 \quad \forall\omega\in\mathbb{R}^d,$$

where K_v: modified Bessel function of the second kind of order v, Γ: Gamma function.

- For $v = \frac{1}{2}$: one gets $k(x,y) = e^{-\frac{\|x-y\|_2}{\sigma}}$.
- Gaussian kernel: $v \to \infty$.

For Poisson kernel: $\sigma \in (0, 1)$.

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp(\widehat{k_0})$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\|x\|}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\|j\|}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm 1,\pm 2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$	$\{0,\pm 1,\pm 2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

Shift-invariant kernels on $\mathbb{R}$ [Sriperumbudur et al., 2010b]

For Poisson kernel: $\sigma \in (0, 1)$.

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp(\widehat{k_0})$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\|x\|}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma, \sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\|j\|}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0, \pm 1, \pm 2, \ldots, \pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$	$\{0, \pm 1, \pm 2, \ldots, \pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma, \sigma\}$

For $x \in \mathbb{R}^d$: $k_0(x) = \prod_{j=1}^d k_0(x_j)$, $\widehat{k_0}(\omega) = \prod_{j=1}^d \widehat{k_0}(\omega_j)$.

B-spline kernel type kernels

- Still k: continuous, bounded, shift-invariant.
- B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.

B-spline kernel type kernels

- Still k: continuous, bounded, shift-invariant.
- B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.
- More generally

Theorem ([Sriperumbudur et al., 2010b])

$supp(k_0)$: compact $\Rightarrow$ k is characteristic.

Construction of new characteristic kernels

Theorem ([Sriperumbudur et al., 2010b])

If k, k_1, k_2: continuous, bounded, shift-invariant; k: characteristic, $k_2 \neq 0$. Then $k + k_1$, kk_2 is also characteristic.

Construction of new characteristic kernels

Theorem ([Sriperumbudur et al., 2010b])

If k, k_1, k_2: continuous, bounded, shift-invariant; k: characteristic, $k_2 \neq 0$. Then $k + k_1$, kk_2 is also characteristic.

Proof.

We focus on $k + k_1$ (product: similarly):

$$(k + k_1)(x, y) := k(x, y) + k_1(x, y) = k_0(x - y) + (k_1)_0(x - y)$$
$$= \int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle} \mathrm{d}(\Lambda + \Lambda_1)(\omega).$$

□

Construction of new characteristic kernels

Theorem ([Sriperumbudur et al., 2010b])

If k, k_1, k_2: continuous, bounded, shift-invariant; k: characteristic, $k_2 \neq 0$. Then $k + k_1$, kk_2 is also characteristic.

Proof.

We focus on $k + k_1$ (product: similarly):

$$(k + k_1)(x, y) := k(x, y) + k_1(x, y) = k_0(x - y) + (k_1)_0(x - y)$$
$$= \int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle} \mathrm{d}(\Lambda + \Lambda_1)(\omega).$$

- k: characteristic $\Rightarrow supp(\Lambda) = \mathbb{R}^d$.

□

Construction of new characteristic kernels

Theorem ([Sriperumbudur et al., 2010b])

If k, k_1, k_2: continuous, bounded, shift-invariant; k: characteristic, $k_2 \neq 0$. Then $k + k_1$, kk_2 is also characteristic.

Proof.

We focus on $k + k_1$ (product: similarly):

$$(k + k_1)(x, y) := k(x, y) + k_1(x, y) = k_0(x - y) + (k_1)_0(x - y)$$
$$= \int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle} \mathrm{d}(\Lambda + \Lambda_1)(\omega).$$

- k: characteristic $\Rightarrow supp(\Lambda) = \mathbb{R}^d$.
- Since $supp(\Lambda) \subseteq supp(\Lambda + \Lambda_1)$, we get $supp(\Lambda + \Lambda_1) = \mathbb{R}^d$; hence $k + k_1$ is characteristic.

□

Radial, bounded, continuous kernels on $\mathbb{R}^d$

Recall (radial kernel):

$$k(x,y) = k_0(\|x-y\|_2), \qquad k_0(z) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2} \mathrm{d}\nu(t).$$

Radial, bounded, continuous kernels on $\mathbb{R}^d$

Recall (radial kernel):

$$k(x,y) = k_0(\|x-y\|_2), \qquad k_0(z) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2} \mathrm{d}\nu(t).$$

Theorem ([Sriperumbudur et al., 2010b])

k is characteristic iff. $supp(\nu) \neq \{0\}$.

More general spaces

- $\mathcal{M}_b(\mathcal{X})$: set of all finite signed (Radon) measures on $\mathcal{X}$ (Radon $\Rightarrow \exists supp$).

More general spaces

- $\mathcal{M}_b(\mathcal{X})$: set of all finite signed (Radon) measures on $\mathcal{X}$ (Radon $\Rightarrow \exists supp$).
- Ulam's Theorem [Dudley, 2004]: On an $\mathcal{X}$ Polish space $\forall$ Borel measure is Radon.

More general spaces

- $\mathcal{M}_b(\mathcal{X})$: set of all finite signed (Radon) measures on $\mathcal{X}$ (Radon $\Rightarrow \exists supp$).
- Ulam's Theorem [Dudley, 2004]: On an $\mathcal{X}$ Polish space $\forall$ Borel measure is Radon.

Definition

A $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ bounded, measurable kernel is called *integrally strictly positive definite (ispd)* if

$$\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \mathrm{d}\mathbb{F}(x)\mathbb{F}(y) > 0 \quad \forall 0 \neq \mathbb{F} \in \mathcal{M}_b(\mathcal{X}).$$

Sufficient condition: ispd

Theorem ([Sriperumbudur et al., 2010b])

Ispd kernels are characteristic on an $\mathcal{X}$ topological space.

- ispd on $\mathbb{R}^d$: Gaussian, Laplacian, inverse multiquadrics, Matérn kernels, B-splines.

Sufficient condition: ispd

Theorem ([Sriperumbudur et al., 2010b])

Ispd kernels are characteristic on an $\mathcal{X}$ topological space.

- ispd on $\mathbb{R}^d$: Gaussian, Laplacian, inverse multiquadrics, Matérn kernels, B-splines.
- Dirichlet kernel: characteristic, though not ispd.

Sufficient condition: ispd

Theorem ([Sriperumbudur et al., 2010b])

Ispd kernels are characteristic on an $\mathcal{X}$ topological space.

- ispd on $\mathbb{R}^d$: Gaussian, Laplacian, inverse multiquadrics, Matérn kernels, B-splines.
- Dirichlet kernel: characteristic, though not ispd.
- ispd property: checking might not be easy.

Ispd: constructions

Shift-variant ispd from shift-invariant ispd kernel:

$$k_0(x,y) = f(x)k(x,y)f(y), \quad f \in C_b(\mathcal{X}).$$

Ispd: constructions

Shift-variant ispd from shift-invariant ispd kernel:

$$k_0(x, y) = f(x)k(x, y)f(y), \quad f \in C_b(\mathcal{X}).$$

Example (exponential ← Gaussian): $k_0(x, y) = e^{\sigma\langle x,y\rangle}$, $\mathcal{X} \subset \mathbb{R}^d$ compact

$$k(x, y) = e^{-\sigma\frac{\|x-y\|^2}{2}}, \qquad f(x) = e^{\sigma\frac{\|x\|^2}{2}}.$$

Denseness in L^r

Theorem ([Fukumizu et al., 2008, Fukumizu et al., 2009a])

Let $r \geqslant 1$.

- *Sufficient condition: A* $k : (\mathcal{X}, \mathcal{A}) \times (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$ *bounded measurable kernel is characteristic if* $\mathcal{H}_k + \mathbb{R}$ *is dense in* $L^r(\mathcal{X}, \mathcal{A}, \mathbb{P})$ *for all* $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$.

Denseness in L^r

Theorem ([Fukumizu et al., 2008, Fukumizu et al., 2009a])

Let $r \geqslant 1$.

- *Sufficient condition: A $k : (\mathcal{X}, \mathcal{A}) \times (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$ bounded measurable kernel is characteristic if $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(\mathcal{X}, \mathcal{A}, \mathbb{P})$ for all $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$.*
- *With $r = 2$, it is also a necessary condition.*

Denseness in L^r

Theorem ([Fukumizu et al., 2008, Fukumizu et al., 2009a])

Let $r \geqslant 1$.

- *Sufficient condition: A $k : (\mathcal{X}, \mathcal{A}) \times (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$ bounded measurable kernel is characteristic if $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(\mathcal{X}, \mathcal{A}, \mathbb{P})$ for all $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$.*
- *With $r = 2$, it is also a necessary condition.*

Note:

- For a c-universal kernel k: sufficient condition holds with $r = 2$.
- This gives the 3rd 'universal $\Rightarrow$ characteristic' proof.

Denseness in L^r

Theorem ([Fukumizu et al., 2008, Fukumizu et al., 2009a])

Let $r \geqslant 1$.

- *Sufficient condition: A $k : (\mathcal{X}, \mathcal{A}) \times (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$ bounded measurable kernel is characteristic if $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(\mathcal{X}, \mathcal{A}, \mathbb{P})$ for all $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$.*
- *With $r = 2$, it is also a necessary condition.*

Note:

- For a c-universal kernel k: sufficient condition holds with $r = 2$.
- This gives the 3rd 'universal $\Rightarrow$ characteristic' proof.

Let us prove this theorem...

Denseness is sufficient: idea

- Goal: in this case, $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P}(A) = \mathbb{Q}(A)$ for any $A \in \mathcal{A}$.

Denseness is sufficient: idea

- Goal: in this case, $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \Rightarrow \mathbb{P}(A) = \mathbb{Q}(A)$ for any $A \in \mathcal{A}$.
- Enough: $|\mathbb{P}(A) - \mathbb{Q}(A)| = |\mathbb{P}\chi_A - \mathbb{Q}\chi_A| \leqslant \epsilon$, $\forall A \in \mathcal{A}$, $\forall \epsilon > 0$.

Denseness is sufficient: idea

- Enough: $|\mathbb{P}(A) - \mathbb{Q}(A)| = |\mathbb{P}\chi_A - \mathbb{Q}\chi_A| \leqslant \epsilon$, $\forall A \in \mathcal{A}$, $\forall \epsilon > 0$.
- Idea:

  using the max. difference of $\mathbb{P}$ and $\mathbb{Q}$ $\Rightarrow$ TV of $\mathbb{P} - \mathbb{Q}$,

$$|\mathbb{P} - \mathbb{Q}|(\mathcal{X}) = 2 \sup_{A \in \mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

Denseness is sufficient: idea

- Enough: $|\mathbb{P}(A) - \mathbb{Q}(A)| = |\mathbb{P}\chi_A - \mathbb{Q}\chi_A| \leq \epsilon$, $\forall A \in \mathcal{A}$, $\forall \epsilon > 0$.
- Idea:

using the max. difference of $\mathbb{P}$ and $\mathbb{Q}$ $\Rightarrow$ TV of $\mathbb{P} - \mathbb{Q}$,

$$|\mathbb{P} - \mathbb{Q}|(\mathcal{X}) = 2 \sup_{A \in \mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

exploit denseness for $\chi_A \in \underbrace{L^r(\mathcal{X}, \mathcal{A}, |\mathbb{P} - \mathbb{Q}|)}_{=: L^r(|\mathbb{P}-\mathbb{Q}|)}$.

Total variation: quick summary

Idea: $f = f^+ - f^- \to |f| = f^+ + f^-$.

Total variation: quick summary

Idea: $f = f^+ - f^- \to |f| = f^+ + f^-$. Analogously:

- $(\mathcal{X}, \mathcal{A})$: measurable space. μ: signed measure on it.

Total variation: quick summary

Idea: $f = f^+ - f^- \to |f| = f^+ + f^-$. Analogously:

- $(\mathcal{X}, \mathcal{A})$: measurable space. μ: signed measure on it.
- Hahn-Jordan decomposition of μ: $\mathcal{X} = \mathcal{P} \dot{\cup} \mathcal{N}$.

Total variation: quick summary

Idea: $f = f^+ - f^- \rightarrow |f| = f^+ + f^-$. Analogously:

- $(\mathcal{X}, \mathcal{A})$: measurable space. μ: signed measure on it.
- Hahn-Jordan decomposition of μ: $\mathcal{X} = \mathcal{P} \,\dot\cup\, \mathcal{N}$.
- Positive & negative part of μ $(= \mu^+ - \mu^-)$:

$$\mu^+(A) = \mu(A \cap \mathcal{P}), \qquad \mu^+(A) = \mu(A \cap \mathcal{N}) \quad \forall A \in \mathcal{A}.$$

Total variation: quick summary

Idea: $f = f^+ - f^- \to |f| = f^+ + f^-$. Analogously:

- $(\mathcal{X}, \mathcal{A})$: measurable space. μ: signed measure on it.
- Hahn-Jordan decomposition of μ: $\mathcal{X} = \mathcal{P} \dot\cup \mathcal{N}$.
- Positive & negative part of μ $(= \mu^+ - \mu^-)$:

$$\mu^+(A) = \mu(A \cap \mathcal{P}), \qquad \mu^+(A) = \mu(A \cap \mathcal{N}) \quad \forall A \in \mathcal{A}.$$

- TV of μ: $|\mu| := \mu^+ + \mu^-$.

Total variation: quick summary

Idea: $f = f^+ - f^- \rightarrow |f| = f^+ + f^-$. Analogously:

- $(\mathcal{X}, \mathcal{A})$: measurable space. μ: signed measure on it.
- Hahn-Jordan decomposition of μ: $\mathcal{X} = \mathcal{P} \dot\cup \mathcal{N}$.
- Positive & negative part of μ $(= \mu^+ - \mu^-)$:

$$\mu^+(A) = \mu(A \cap \mathcal{P}), \qquad \mu^+(A) = \mu(A \cap \mathcal{N}) \quad \forall A \in \mathcal{A}.$$

- TV of μ: $|\mu| := \mu^+ + \mu^-$.
- μ: finite $\Rightarrow \mu^+$, μ^-: finite.

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

- Some lower bounding

$$\epsilon \geqslant \|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \overset{r \geqslant 1}{\gtrsim} \|\underbrace{f - \chi_A}_{=:g}\|_{L^1(|\mathbb{P}-\mathbb{Q}|)}$$

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

- Some lower bounding

$$\epsilon \geqslant \|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \overset{r\geqslant 1}{\gtrsim} \|\underbrace{f - \chi_A}_{=:g}\|_{L^1(|\mathbb{P}-\mathbb{Q}|)} = |\mathbb{P} - \mathbb{Q}|(|g|)$$

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

- Some lower bounding

$$\epsilon \geqslant \|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \overset{r\geqslant 1}{\gtrsim} \|\underbrace{f - \chi_A}_{=:g}\|_{L^1(|\mathbb{P}-\mathbb{Q}|)} = |\mathbb{P} - \mathbb{Q}|(|g|)$$

$$\geqslant |\mathbb{P} - \mathbb{Q}|(g)$$

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

- Some lower bounding

$$\epsilon \geqslant \|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \overset{r \geqslant 1}{\gtrsim} \|\underbrace{f - \chi_A}_{=:g}\|_{L^1(|\mathbb{P}-\mathbb{Q}|)} = |\mathbb{P} - \mathbb{Q}|(|g|)$$

$$\geqslant |\mathbb{P} - \mathbb{Q}|(g) \geqslant |(\mathbb{P} - \mathbb{Q})(g)|$$

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

- Some lower bounding

$$\epsilon \geqslant \|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \overset{r \geqslant 1}{\gtrsim} \|\underbrace{f - \chi_A}_{=:g}\|_{L^1(|\mathbb{P}-\mathbb{Q}|)} = |\mathbb{P} - \mathbb{Q}|(|g|)$$

$$\geqslant |\mathbb{P} - \mathbb{Q}|(g) \geqslant |(\mathbb{P} - \mathbb{Q})(g)| = |\mathbb{P}(f - \chi_A) - \mathbb{Q}(f - \chi_A)|$$

Denseness is sufficient: proof

- Take: $A \in \mathcal{A}$, $\epsilon > 0$.
- $\mathcal{H}_k + \mathbb{R}$ is dense in $L^r(|\mathbb{P} - \mathbb{Q}|) \Rightarrow \exists f \in \mathcal{H}_k + \mathbb{R}$

$$\|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \leqslant \epsilon.$$

- Some lower bounding

$$\epsilon \geqslant \|f - \chi_A\|_{L^r(|\mathbb{P}-\mathbb{Q}|)} \overset{r\geqslant 1}{\gtrsim} \|\underbrace{f - \chi_A}_{=:g}\|_{L^1(|\mathbb{P}-\mathbb{Q}|)} = |\mathbb{P} - \mathbb{Q}|(|g|)$$

$$\geqslant |\mathbb{P} - \mathbb{Q}|(g) \geqslant |(\mathbb{P} - \mathbb{Q})(g)| = |\mathbb{P}(f - \chi_A) - \mathbb{Q}(f - \chi_A)|$$

$$\overset{(*)}{=} |\mathbb{P}\chi_A - \mathbb{Q}\chi_A|.$$

(*): $\mathbb{P}f = \mathbb{Q}f$ for any $f \in \mathcal{H}_k$ since $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$.

Denseness in L^2 is necessary: proof

If $\mathcal{H}_k + \mathbb{R}$ is *not dense* in $L^2(\mathbb{P}) := L^2(\mathcal{X}, \mathcal{A}, \mathbb{P})$, then

- goal: $\underbrace{\exists \mathbb{Q}_1 \neq \mathbb{Q}_2 \in \mathcal{M}_1^+(\mathcal{X}) \text{ st. } \mu_{\mathbb{Q}_1} = \mu_{\mathbb{Q}_2}}_{\mu \text{ is not injective}}$.

Denseness in L^2 is necessary: proof

If $\mathcal{H}_k + \mathbb{R}$ is *not* dense in $L^2(\mathbb{P}) := L^2(\mathcal{X}, \mathcal{A}, \mathbb{P})$, then

- goal: $\underbrace{\exists \mathbb{Q}_1 \ne \mathbb{Q}_2 \in \mathcal{M}_1^+(\mathcal{X}) \text{ st. } \mu_{\mathbb{Q}_1} = \mu_{\mathbb{Q}_2}}_{\mu \text{ is not injective}}$.
- Hahn-Banach: $0 \ne f \in L^2(\mathbb{P})$ st. $f \perp 1, \mathcal{H}_k$, thus

$$\langle f, 1 \rangle_{L^2(\mathbb{P})} = 0, \qquad \langle f, h \rangle_{L^2(\mathbb{P})} = 0 \quad (\forall h \in \mathcal{H}_k).$$

Denseness in L^2 is necessary: proof

If $\mathcal{H}_k + \mathbb{R}$ is *not dense* in $L^2(\mathbb{P}) := L^2(\mathcal{X}, \mathcal{A}, \mathbb{P})$, then

- goal: $\underbrace{\exists \mathbb{Q}_1 \neq \mathbb{Q}_2 \in \mathcal{M}_1^+(\mathcal{X}) \text{ st. } \mu_{\mathbb{Q}_1} = \mu_{\mathbb{Q}_2}}_{\mu \text{ is not injective}}$.
- Hahn-Banach: $0 \neq f \in L^2(\mathbb{P})$ st. $f \perp 1, \mathcal{H}_k$, thus

$$\langle f, 1 \rangle_{L^2(\mathbb{P})} = 0, \qquad \langle f, h \rangle_{L^2(\mathbb{P})} = 0 \quad (\forall h \in \mathcal{H}_k).$$

- We define $\mathbb{Q}_1$, $\mathbb{Q}_2 \in \mathcal{M}_1^+(\mathcal{X})$ from f ($f \neq 0 \Rightarrow \mathbb{Q}_1 \neq \mathbb{Q}_2$):

$$\mathbb{Q}_1(A) = c \int_A |f| \mathrm{d}\mathbb{P}, \quad \mathbb{Q}_2(A) = c \int_A (\underbrace{|f| - f}_{\geqslant 0}) \mathrm{d}\mathbb{P}, \quad c = \frac{1}{\int_{\mathcal{X}} |f| \mathrm{d}\mathbb{P}}.$$

Denseness in L^2 is necessary: proof continued

We arrive at

$$\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} = \int k(\cdot, x) \mathrm{d}\mathbb{Q}_1(x) - \int k(\cdot, x) \mathrm{d}\mathbb{Q}_2(x)$$

Denseness in L^2 is necessary: proof continued

We arrive at

$$\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} = \int k(\cdot, x)\mathrm{d}\mathbb{Q}_1(x) - \int k(\cdot, x)\mathrm{d}\mathbb{Q}_2(x)$$
$$= \int k(\cdot, x)\mathrm{d}(\mathbb{Q}_1 - \mathbb{Q}_2)(x)$$

Denseness in L^2 is necessary: proof continued

We arrive at

$$\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} = \int k(\cdot, x)\mathrm{d}\mathbb{Q}_1(x) - \int k(\cdot, x)\mathrm{d}\mathbb{Q}_2(x)$$
$$= \int k(\cdot, x)\mathrm{d}(\mathbb{Q}_1 - \mathbb{Q}_2)(x) = c\int_{\mathcal{X}} f(x)k(\cdot, x)\mathrm{d}\mathbb{P}(x),$$

Denseness in L^2 is necessary: proof continued

We arrive at

$$\begin{aligned}\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} &= \int k(\cdot, x)\mathrm{d}\mathbb{Q}_1(x) - \int k(\cdot, x)\mathrm{d}\mathbb{Q}_2(x)\\ &= \int k(\cdot, x)\mathrm{d}(\mathbb{Q}_1 - \mathbb{Q}_2)(x) = c\int_{\mathcal{X}} f(x)k(\cdot, x)\mathrm{d}\mathbb{P}(x),\end{aligned}$$

$$(\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2})(y) = c\int_{\mathcal{X}} f(x)k(y, x)\mathrm{d}\mathbb{P}(x)$$

We arrive at

$$\begin{aligned}
\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} &= \int k(\cdot, x)\mathrm{d}\mathbb{Q}_1(x) - \int k(\cdot, x)\mathrm{d}\mathbb{Q}_2(x) \\
&= \int k(\cdot, x)\mathrm{d}(\mathbb{Q}_1 - \mathbb{Q}_2)(x) = c\int_{\mathcal{X}} f(x)k(\cdot, x)\mathrm{d}\mathbb{P}(x), \\
(\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2})(y) &= c\int_{\mathcal{X}} f(x)k(y, x)\mathrm{d}\mathbb{P}(x) \\
&= c\langle f, \underbrace{k(y, \cdot)}_{\in \mathcal{H}_k} \rangle_{L^2(\mathbb{P})} = 0 \quad (\forall y \in \mathcal{X}).
\end{aligned}$$

Denseness in L^2 is necessary: proof continued

We arrive at

$$\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} = \int k(\cdot, x)\mathrm{d}\mathbb{Q}_1(x) - \int k(\cdot, x)\mathrm{d}\mathbb{Q}_2(x)$$
$$= \int k(\cdot, x)\mathrm{d}(\mathbb{Q}_1 - \mathbb{Q}_2)(x) = c\int_{\mathcal{X}} f(x)k(\cdot, x)\mathrm{d}\mathbb{P}(x),$$
$$(\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2})(y) = c\int_{\mathcal{X}} f(x)k(y, x)\mathrm{d}\mathbb{P}(x)$$
$$= c\langle f, \underbrace{k(y, \cdot)}_{\in \mathcal{H}_k} \rangle_{L^2(\mathbb{P})} = 0 \quad (\forall y \in \mathcal{X}).$$

Thus $\mu_{\mathbb{Q}_1} - \mu_{\mathbb{Q}_2} = 0$ despite $\mathbb{Q}_1 \neq \mathbb{Q}_2$.

Infinitely divisible distributions: quick summary

U: random variable.

Question

Can it be decomposed to the sum of 2 i.i.d. random variables?

Infinitely divisible distributions: quick summary

U: random variable.

Question

Can it be decomposed to the sum of 2 i.i.d. random variables?

Question

Can it be decomposed to the sum of 3 i.i.d. random variables?

Infinitely divisible distributions: quick summary

U: random variable.

Question

Can it be decomposed to the sum of 2 i.i.d. random variables?

Question

Can it be decomposed to the sum of 3 i.i.d. random variables?

Question

Can it be decomposed to the sum of 4 i.i.d. random variables?

Infinitely divisible distributions: quick summary

U: random variable.

Question

Can it be decomposed to the sum of 2 i.i.d. random variables?

Question

Can it be decomposed to the sum of 3 i.i.d. random variables?

Question

Can it be decomposed to the sum of 4 i.i.d. random variables?

Question

Can it be decomposed to the sum of n i.i.d. random variables for any $n \in \mathbb{Z}^+$?

Infinitely divisible distributions: quick summary

Examples:

- Poisson, negative binomial, Gamma distribution, student t.

Infinitely divisible distributions: quick summary

Examples:

- Poisson, negative binomial, Gamma distribution, student t.
- normal, Cauchy distribution

Infinitely divisible distributions: quick summary

Examples:

- Poisson, negative binomial, Gamma distribution, student t.
- normal $(\alpha = 2)$, Cauchy distribution $(\alpha = 1)$ $\xleftarrow{\text{spec.}}$ $\forall$ α-stable.

Infinitely divisible distributions: quick summary

Examples:

- Poisson, negative binomial, Gamma distribution, student t.
- normal $(\alpha = 2)$, Cauchy distribution $(\alpha = 1)$ $\xleftarrow{\text{spec.}}$ $\forall$ α-stable.

Counterexamples:

- uniform, binomial distribution

Infinitely divisible distributions: quick summary

Examples:

- Poisson, negative binomial, Gamma distribution, student t.
- normal ($\alpha = 2$), Cauchy distribution ($\alpha = 1$) $\xleftarrow{\text{spec.}}$ $\forall$ α-stable.

Counterexamples:

- uniform, binomial distribution $\xleftarrow{\text{spec.}}$ $\forall$ any distribution with bounded (finite) support.

Symmetric infinitely divisible on $\mathbb{R}^d \Rightarrow$ characteristic

Theorem ([Nishiyama and Fukumizu, 2016])

Assume

- $k(x, y) = k_0(x - y)$, $k_0 \in C_b(\mathbb{R}^d)$, k_0 *is the pdf of*
- *an infinitely divisible, symmetric distribution.*

Then k *is characteristic.*

Symmetric infinitely divisible on $\mathbb{R}^d \Rightarrow$ characteristic

Theorem ([Nishiyama and Fukumizu, 2016])

Assume

- $k(x, y) = k_0(x - y)$, $k_0 \in C_b(\mathbb{R}^d)$, k_0 *is the pdf of*
- *an infinitely divisible, symmetric distribution.*

Then k *is characteristic.*

Examples: Gaussian, Matérn kernel, α-stable kernels, student t-kernels, ...

Characteristic kernels: finished.

Local summary

- Dependency measure applications.
- KCCA. Mean embedding: $\mu_{\mathbb{P}} = \int_X k(\cdot, x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k$.
- Injectivity of μ on
  - probability distributions: characteristic property.
  - finite signed measures: universality ($\mathcal{X}$: compact metric).
- By definition: injectivity of μ $\Leftrightarrow$

$$d_k(\mathbb{P}, \mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}$$

  is a metric.

Maximum mean discrepancy (MMD)

MMD is a specific integral probability metric (IPM)

- $\mathcal{F} = \left\{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} = 1 \right\}$: unit ball in $\mathcal{H}_k$.

$$d_k(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}$$

MMD is a specific integral probability metric (IPM)

- $\mathcal{F} = \left\{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} = 1 \right\}$: unit ball in $\mathcal{H}_k$.

$$d_k(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}$$
$$= \sup_{f \in \mathcal{F}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$$

MMD is a specific integral probability metric (IPM)

- $\mathcal{F} = \left\{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} = 1 \right\}$: unit ball in $\mathcal{H}_k$.

$$\begin{aligned} d_k(\mathbb{P},\mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} \\ &= \sup_{f\in\mathcal{F}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} \\ &= \sup_{f\in\mathcal{F}} (\mathbb{P}f - \mathbb{Q}f). \end{aligned}$$

MMD is a specific integral probability metric (IPM)

- $\mathcal{F} = \left\{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} = 1 \right\}$: unit ball in $\mathcal{H}_k$.

$$d_k(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{F}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{Q}f).$$

- IPMs [Zolotarev, 1983, Müller, 1997].

IPM: other $\mathcal{F}$ examples giving metric

- $\mathcal{F} = C_b(\mathcal{X})$ with $\mathcal{X}$ metric space.

IPM: other $\mathcal{F}$ examples giving metric

- $\mathcal{F} = C_b(\mathcal{X})$ with $\mathcal{X}$ metric space.
- $\mathcal{F} = \{f : \|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)| \leqslant 1\}$:
  - bounded functions.
  - total variation distance.

IPM: other $\mathcal{F}$ examples giving metric

- $\mathcal{F} = C_b(\mathcal{X})$ with $\mathcal{X}$ metric space.
- $\mathcal{F} = \{f : \|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)| \leqslant 1\}$:
  - bounded functions.
  - total variation distance.
- $\mathcal{F} = \left\{f : \|f\|_L := \sup_{x\neq y} \frac{|f(x)-f(y)|}{\rho(x,y)} \leqslant 1\right\}$:
  - Kantorovich metric $\xrightarrow{\mathcal{X}\text{: separable metric}}$ Wasserstein distance.

IPM: other $\mathcal{F}$ examples giving metric

- $\mathcal{F} = \{f : \|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)| \leqslant 1\}$:
  - bounded functions.
  - total variation distance.

TV upper bounds MMD [Sriperumbudur et al., 2010b]:

$$d_k(\mathbb{P},\mathbb{Q}) \leqslant \sup_{x\in\mathcal{X}} \sqrt{k(x,x)}\, TV(\mathbb{P},\mathbb{Q}).$$

IPM: other $\mathcal{F}$ examples giving metric – continued

- $\mathcal{F} = \{f : \|f\|_{BL} := \|f\|_\infty + \|f\|_L \leqslant 1\}$
  - bounded Lipschitz functions,
  - Dudley metric.

IPM: other $\mathcal{F}$ examples giving metric – continued

- $\mathcal{F} = \{f : \|f\|_{BL} := \|f\|_\infty + \|f\|_L \leqslant 1\}$
  - bounded Lipschitz functions,
  - Dudley metric.
- $\mathcal{F} = \left\{\chi_{(-\infty,t]} : t \in \mathbb{R}^d\right\}$:
  - characteristic functions of half-intervals.
  - Kolmogorov distance.

Empirical estimation of IPMs

[Sriperumbudur et al., 2012]:

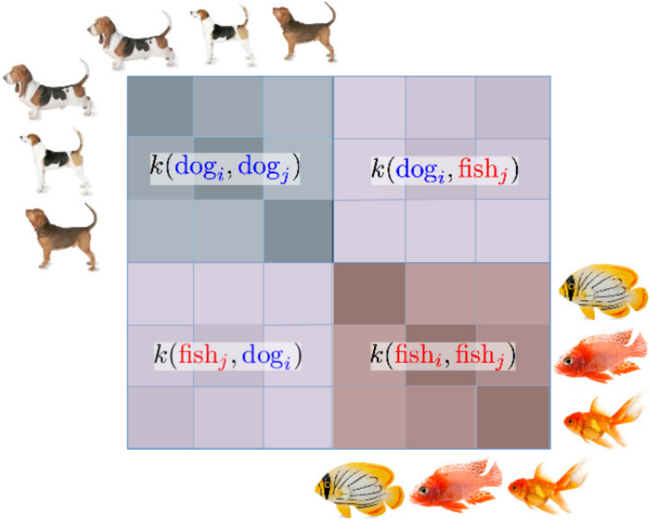
- Kantorovich, Dudley metric: linear programming task.
- MMD (d_k): easier.

MMD estimators

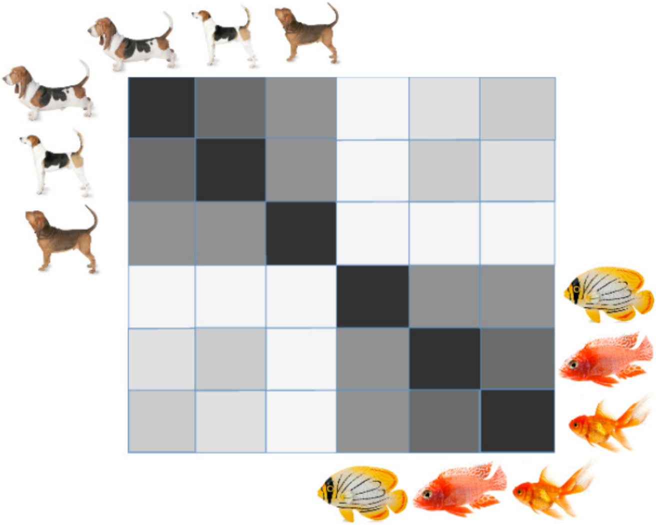
$\sim P$

$\sim Q$

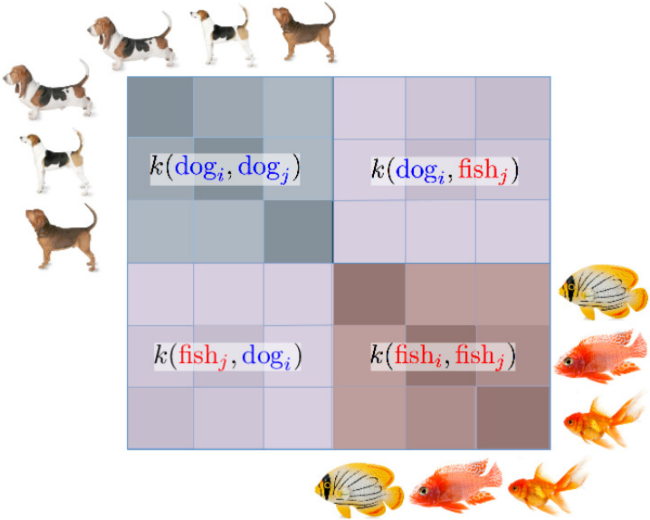
MMD estimator: intuition



MMD estimator: intuition



MMD estimator: intuition



$$\widehat{MMD}^2(\mathbb{P}, \mathbb{Q}) = \overline{G_{\mathbb{P},\mathbb{P}}} + \overline{G_{\mathbb{Q},\mathbb{Q}}} - 2\overline{G_{\mathbb{P},\mathbb{Q}}} \text{ (without diagonals in } \overline{G_{\mathbb{P},\mathbb{P}}}, \overline{G_{\mathbb{Q},\mathbb{Q}}})$$

† $\widehat{MMD}$ & $\widehat{HSIC}$ illustration credit: Arthur Gretton

MMD estimator-1

Recall: MMD = squared difference between feature means:

$$
\begin{aligned}
MMD^2(\mathbb{P},\mathbb{Q}) := d_k^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}^2 = \\
= \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}} k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}} k(y,y') \\
- 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}} k(x,y).
\end{aligned}
$$

MMD estimator-1

Recall: MMD = squared difference between feature means:

$$MMD^2(\mathbb{P},\mathbb{Q}) := d_k^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}^2 = $$
$$= \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}} k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}} k(y,y') - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}} k(x,y).$$

Unbiased empirical estimator using $\{x_i\}_{i=1}^m \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$:

$$\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q}) = \overline{G_{\mathbb{P},\mathbb{P}}} + \overline{G_{\mathbb{Q},\mathbb{Q}}} - 2\overline{G_{\mathbb{P},\mathbb{Q}}}$$

MMD estimator-1

Recall: MMD = squared difference between feature means:

$$MMD^2(\mathbb{P},\mathbb{Q}) := d_k^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k} = \\ = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}} k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}} k(y,y') \\ - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}} k(x,y).$$

Unbiased empirical estimator using $\{x_i\}_{i=1}^m \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$:

$$\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q}) = \overline{G_{\mathbb{P},\mathbb{P}}} + \overline{G_{\mathbb{Q},\mathbb{Q}}} - 2\overline{G_{\mathbb{P},\mathbb{Q}}} \\ = \underbrace{\frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(x_i,x_j)}_{\text{U-statistic-1}} + \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i} k(y_i,y_j)}_{\text{U-statistic-2}} \\ - \underbrace{\frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i,y_j)}_{\text{sample average}}.$$

MMD estimator-2

We plug in the empirical measures $(\mathbb{P}_m, \mathbb{Q}_n)$:

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k},$$
$$\widehat{MMD}^2_b(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}_m} - \mu_{\mathbb{Q}_n}\|^2_{\mathcal{H}_k}$$

MMD estimator-2

We plug in the empirical measures $(\mathbb{P}_m, \mathbb{Q}_n)$:

$$
\begin{aligned}
MMD^2(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k}, \\
\widehat{MMD}^2_b(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}_m} - \mu_{\mathbb{Q}_n}\|^2_{\mathcal{H}_k} \\
&= \|\mu_{\mathbb{P}_m}\|^2_{\mathcal{H}_k} + \|\mu_{\mathbb{Q}_n}\|^2_{\mathcal{H}_k} - 2\langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{Q}_n} \rangle_{\mathcal{H}_k}.
\end{aligned}
$$

MMD estimator-2

We plug in the empirical measures ($\mathbb{P}_m$, $\mathbb{Q}_n$):

$$
\begin{aligned}
MMD^2(\mathbb{P},\mathbb{Q}) &= \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}^2, \\
\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) &= \left\|\mu_{\mathbb{P}_m} - \mu_{\mathbb{Q}_n}\right\|_{\mathcal{H}_k}^2 \\
&= \left\|\mu_{\mathbb{P}_m}\right\|_{\mathcal{H}_k}^2 + \left\|\mu_{\mathbb{Q}_n}\right\|_{\mathcal{H}_k}^2 - 2\left\langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{Q}_n}\right\rangle_{\mathcal{H}_k}.
\end{aligned}
$$

Enough:

$$
\left\langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{Q}_n}\right\rangle_{\mathcal{H}_k} = \left\langle \frac{1}{m}\sum_{i=1}^{m} k(\cdot, x_i), \frac{1}{n}\sum_{j=1}^{n} k(\cdot, y_j)\right\rangle_{\mathcal{H}_k}
$$

MMD estimator-2

We plug in the empirical measures $(\mathbb{P}_m, \mathbb{Q}_n)$:

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k},$$
$$\widehat{MMD}^2_b(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}_m} - \mu_{\mathbb{Q}_n}\|^2_{\mathcal{H}_k}$$
$$= \|\mu_{\mathbb{P}_m}\|^2_{\mathcal{H}_k} + \|\mu_{\mathbb{Q}_n}\|^2_{\mathcal{H}_k} - 2\langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{Q}_n}\rangle_{\mathcal{H}_k}.$$

Enough:

$$\langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{Q}_n}\rangle_{\mathcal{H}_k} = \left\langle \frac{1}{m}\sum_{i=1}^{m} k(\cdot, x_i), \frac{1}{n}\sum_{j=1}^{n} k(\cdot, y_j) \right\rangle_{\mathcal{H}_k}$$
$$= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \underbrace{\langle k(\cdot, x_i), k(\cdot, y_j)\rangle_{\mathcal{H}_k}}_{k(x_i, y_j)}.$$

MMD estimator-2: continued

$$\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \underbrace{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i,x_j)}_{\text{V-statistic-1}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i,y_j)}_{\text{V-statistic-2}}$$

$$- \underbrace{\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j)}_{\text{sample average}}.$$

MMD estimator-2: continued

$$\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \underbrace{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i,x_j)}_{\text{V-statistic-1}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(y_i,y_j)}_{\text{V-statistic-2}} - \underbrace{\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j)}_{\text{sample average}}.$$

Notes:

- $\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q})$: unbiased; it might be negative.

MMD estimator-2: continued

$$\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \underbrace{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i,x_j)}_{\text{V-statistic-1}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i,y_j)}_{\text{V-statistic-2}} - \underbrace{\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j)}_{\text{sample average}}.$$

Notes:

- $\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q})$: unbiased; it might be negative.
- $\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}_m} - \mu_{\mathbb{Q}_n}\|_{\mathcal{H}_k}^2 \geqslant 0$.

MMD estimator-2: continued

$$\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \underbrace{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i,x_j)}_{\text{V-statistic-1}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(y_i,y_j)}_{\text{V-statistic-2}}$$

$$- \underbrace{\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j)}_{\text{sample average}}.$$

Notes:

- $\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q})$: unbiased; it might be negative.
- $\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}_m} - \mu_{\mathbb{Q}_n}\|_{\mathcal{H}_k}^2 \geqslant 0$.
- Computational complexity: $\mathcal{O}\left((m+n)^2\right)$, quadratic.

Notes

- Set kernel, convolution kernel.

Notes

- Set kernel, convolution kernel.
- Other valid $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$ examples $\rightarrow$ distribution classification [Póczos et al., 2012, Muandet et al., 2011] / distribution regression [Szabó et al., 2016].

Notes

- Set kernel, convolution kernel.
- Other valid $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$ examples → distribution classification [Póczos et al., 2012, Muandet et al., 2011] / distribution regression [Szabó et al., 2016].
- Few analytic expressions exist for MMD.

Notes

- Set kernel, convolution kernel.
- Other valid $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$ examples $\rightarrow$ distribution classification [Póczos et al., 2012, Muandet et al., 2011] / distribution regression [Szabó et al., 2016].
- Few analytic expressions exist for MMD.
- Embedding to RKBS.

Notes

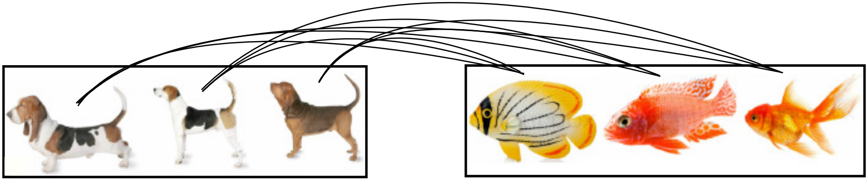
- Set kernel, convolution kernel.
- Other valid $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$ examples $\rightarrow$ distribution classification [Póczos et al., 2012, Muandet et al., 2011] / distribution regression [Szabó et al., 2016].
- Few analytic expressions exist for MMD.
- Embedding to RKBS.

Let us see the details.

Set kernel

Convolution kernels [Haussler, 1999] $\ni$ set kernel [Gärtner et al., 2002]:

$$K(\mathbb{P}_m, \mathbb{Q}_n) := \langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{Q}_n} \rangle_{\mathcal{H}_k} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j).$$



Other valid K examples [Christmann and Steinwart, 2010], [Szabó et al., 2015] → distribution regression

Recall: $K(\mathbb{P},\mathbb{Q}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$, linear kernel.

K_G	K_e	K_C
$e^{-\frac{\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}}{2\theta^2}}$	$e^{-\frac{\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}}{2\theta^2}}$	$\left(1 + \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}^2 / \theta^2\right)^{-1}$

K_t	K_i
$\left(1 + \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}^{\theta}\right)^{-1}$	$\left(\left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}^2 + \theta^2\right)^{-\frac{1}{2}}$

Other valid K examples [Christmann and Steinwart, 2010], [Szabó et al., 2015] → distribution regression

Recall: $K(\mathbb{P},\mathbb{Q}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$, linear kernel.

K_G	K_e	K_C
$e^{-\frac{\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}}{2\theta^2}}$	$e^{-\frac{\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}}{2\theta^2}}$	$\left(1+\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}^2/\theta^2\right)^{-1}$

K_t	K_i
$\left(1+\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}{}^{\theta}\right)^{-1}$	$\left(\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}^2+\theta^2\right)^{-\frac{1}{2}}$

Functions of $\left\|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}$ ⇒ computation: similar to set kernel.

Few analytic expressions exist: examples [Gretton et al., 2007, Muandet et al., 2011]

Assume: $\mathbb{P} = N(m_1, \Sigma_1)$, $\mathbb{Q} = N(m_2, \Sigma_2)$.

$k(x, y)$	$K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$
$e^{-\frac{\gamma}{2}\|x-y\|_2^2}$	$\frac{e^{-\frac{1}{2}(m_1-m_2)^T(\Sigma_1+\Sigma_2+\gamma I)^{-1}(m_1-m_2)}}{\|\gamma\Sigma_1+\gamma\Sigma_2+I\|^{\frac{1}{2}}}$

Few analytic expressions exist: examples [Gretton et al., 2007, Muandet et al., 2011]

Assume: $\mathbb{P} = N(m_1, \Sigma_1)$, $\mathbb{Q} = N(m_2, \Sigma_2)$.

$k(x,y)$	$K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$
$e^{-\frac{\gamma}{2}\|x-y\|_2^2}$	$\frac{e^{-\frac{1}{2}(m_1-m_2)^T(\Sigma_1+\Sigma_2+\gamma I)^{-1}(m_1-m_2)}}{\lvert\gamma\Sigma_1+\gamma\Sigma_2+I\rvert^{\frac{1}{2}}}$
$(1+\langle x,y\rangle)^2$	$(1+\langle m_1, m_2\rangle)^2 + \operatorname{tr}(\Sigma_1\Sigma_2) + m_1\Sigma_2 m_1 + m_2\Sigma_1 m_2$

Few analytic expressions exist: examples [Gretton et al., 2007, Muandet et al., 2011]

Assume: $\mathbb{P} = N(m_1, \Sigma_1)$, $\mathbb{Q} = N(m_2, \Sigma_2)$.

$k(x,y)$	$K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$
$e^{-\frac{\gamma}{2}\|x-y\|_2^2}$	$\frac{e^{-\frac{1}{2}(m_1-m_2)^T(\Sigma_1+\Sigma_2+\gamma I)^{-1}(m_1-m_2)}}{\|\gamma\Sigma_1+\gamma\Sigma_2+I\|^{\frac{1}{2}}}$
$(1+\langle x,y\rangle)^2$	$(1+\langle m_1,m_2\rangle)^2 + \mathrm{tr}(\Sigma_1\Sigma_2) + m_1\Sigma_2 m_1 + m_2\Sigma_1 m_2$
$(1+\langle x,y\rangle)^3$	$(1+\langle m_1,m_2\rangle)^3 + 6m_1^T\Sigma_1\Sigma_2 m_2 + 3(1+\langle m_1,m_2\rangle)\times$ $[\mathrm{tr}(\Sigma_1\Sigma_2) + m_1\Sigma_2 m_1 + m_2\Sigma_1 m_2]$

RKBS definition

Definition ([Zhang et al., 2009])

A reflexive $\mathcal{B}$ *Banach space* [$(\mathcal{B}')' = \mathcal{B}$] *of functions on* $\mathcal{X}$ *is called RKBS if*

RKBS definition

Definition ([Zhang et al., 2009])

A reflexive $\mathcal{B}$ *Banach space* [$(\mathcal{B}')' = \mathcal{B}$] *of functions on* $\mathcal{X}$ *is called RKBS if*

- $\mathcal{B}'$ *is isometric to a Banach space of functions on* $\mathcal{X}$*, and*

RKBS definition

Definition ([Zhang et al., 2009])

A reflexive $\mathcal{B}$ Banach space $[(\mathcal{B}')' = \mathcal{B}]$ of functions on $\mathcal{X}$ is called RKBS if

- *$\mathcal{B}'$ is isometric to a Banach space of functions on $\mathcal{X}$, and*
- *the evaluation functional is continuous on both $\mathcal{B}$ and $\mathcal{B}'$.*

RKBS definition

Definition ([Zhang et al., 2009])

A reflexive $\mathcal{B}$ Banach space $[(\mathcal{B}')' = \mathcal{B}]$ of functions on $\mathcal{X}$ is called RKBS if

- *$\mathcal{B}'$ is isometric to a Banach space of functions on $\mathcal{X}$, and*
- *the evaluation functional is continuous on both $\mathcal{B}$ and $\mathcal{B}'$.*

Notes:

- Generally, $\mathcal{B} \subseteq \mathcal{B}''$.

RKBS definition

Definition ([Zhang et al., 2009])

A reflexive $\mathcal{B}$ Banach space [$(\mathcal{B}')' = \mathcal{B}$] of functions on $\mathcal{X}$ is called RKBS if

- *$\mathcal{B}'$ is isometric to a Banach space of functions on $\mathcal{X}$, and*
- *the evaluation functional is continuous on both $\mathcal{B}$ and $\mathcal{B}'$.*

Notes:

- Generally, $\mathcal{B} \subseteq \mathcal{B}''$.
- For $\mathcal{B} = \mathcal{H}$ Hilbert: $(\mathcal{H}')' = \mathcal{H}$ (Riesz representation theorem).

RKBS properties

Using the

$$\langle f, g' \rangle_{\mathcal{B}} := g'(f), \quad (f \in \mathcal{B}, g' \in \mathcal{B}')$$

notation

RKBS properties

Using the

$$\langle f, g' \rangle_{\mathcal{B}} := g'(f), \quad (f \in \mathcal{B}, g' \in \mathcal{B}')$$

notation

$$k(\cdot, x) \in \mathcal{B}' \quad (\forall x \in \mathcal{X}), \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{B}} \quad \forall f \in \mathcal{B},$$

RKBS properties

Using the

$$\langle f, g' \rangle_{\mathcal{B}} := g'(f), \quad (f \in \mathcal{B}, g' \in \mathcal{B}')$$

notation

$$k(\cdot, x) \in \mathcal{B}' \quad (\forall x \in \mathcal{X}), \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{B}} \quad \forall f \in \mathcal{B},$$

$$k(x, \cdot) \in \mathcal{B} \quad (\forall x \in \mathcal{X}), \qquad f'(x) = \langle k(x, \cdot), f' \rangle_{\mathcal{B}} \quad \forall f' \in \mathcal{B}',$$

RKBS properties

Using the

$$\langle f, g' \rangle_{\mathcal{B}} := g'(f), \quad (f \in \mathcal{B}, g' \in \mathcal{B}')$$

notation

$$k(\cdot, x) \in \mathcal{B}' \quad (\forall x \in \mathcal{X}), \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{B}} \quad \forall f \in \mathcal{B},$$

$$k(x, \cdot) \in \mathcal{B} \quad (\forall x \in \mathcal{X}), \qquad f'(x) = \langle k(x, \cdot), f' \rangle_{\mathcal{B}} \quad \forall f' \in \mathcal{B}',$$

$$\mathcal{B} = \overline{span\left\{k(x, \cdot), x \in \mathcal{X}\right\}}, \qquad \mathcal{B}' = \overline{span\left\{k(\cdot, x), x \in \mathcal{X}\right\}},$$

RKBS properties

Using the

$$\langle f, g' \rangle_{\mathcal{B}} := g'(f), \quad (f \in \mathcal{B}, g' \in \mathcal{B}')$$

notation

$$k(\cdot, x) \in \mathcal{B}' \quad (\forall x \in \mathcal{X}),$$
$$k(x, \cdot) \in \mathcal{B} \quad (\forall x \in \mathcal{X}),$$
$$\mathcal{B} = \overline{span\{k(x, \cdot), x \in \mathcal{X}\}},$$
$$k(x, y) = \langle k(x, \cdot), k(\cdot, y) \rangle_{\mathcal{B}} \, \forall x, y \in \mathcal{X}.$$

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{B}} \quad \forall f \in \mathcal{B},$$
$$f'(x) = \langle k(x, \cdot), f' \rangle_{\mathcal{B}} \quad \forall f' \in \mathcal{B}',$$
$$\mathcal{B}' = \overline{span\{k(\cdot, x), x \in \mathcal{X}\}},$$

Pecularities of RKBS-s

- Different RKBSs can have the same r.k.

Pecularities of RKBS-s

- Different RKBSs can have the same r.k.
- No inner product on $\mathcal{B}$ $\Rightarrow$ an r.k. can be an arbitrary function.

Pecularities of RKBS-s

- Different RKBSs can have the same r.k.
- No inner product on $\mathcal{B} \Rightarrow$ an r.k. can be an arbitrary function.
- For specific RKBSs†:

†Uniformly Fréchet differentiable and uniformly convex.

Pecularities of RKBS-s

- Different RKBSs can have the same r.k.
- No inner product on $\mathcal{B}$ $\Rightarrow$ an r.k. can be an arbitrary function.
- For specific RKBSs[†]:
  - 'Riesz representation theorem' exists,

[†] Uniformly Fréchet differentiable and uniformly convex.

Pecularities of RKBS-s

- Different RKBSs can have the same r.k.
- No inner product on $\mathcal{B}$ $\Rightarrow$ an r.k. can be an arbitrary function.
- For specific RKBSs[†]:
  - 'Riesz representation theorem' exists,
  - $k \leftrightarrow G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.i.p.

[†] Uniformly Fréchet differentiable and uniformly convex.

Pecularities of RKBS-s

- Different RKBSs can have the same r.k.
- No inner product on $\mathcal{B} \Rightarrow$ an r.k. can be an arbitrary function.
- For specific RKBSs[†]:
  - 'Riesz representation theorem' exists,
  - $k \leftrightarrow G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.i.p.
  - $\mu_{\mathbb{P}} = \int_X \underbrace{k(\cdot, x)}_{\in \mathcal{B}'} \mathrm{d}\mathbb{P}(x) \in \mathcal{B}'$ [Sriperumbudur et al., 2011].

[†] Uniformly Fréchet differentiable and uniformly convex.

RKBS: computational intractability

Key for RKHS $\mathcal{H}_k$:

$$d_k^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \mathrm{d}(\mathbb{P} - \mathbb{Q})(x) \mathrm{d}(\mathbb{P} - \mathbb{Q})(y).$$

RKBS: computational intractability

Key for RKHS $\mathcal{H}_k$:

$$d_k{}^2(\mathbb{P},\mathbb{Q}) = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y).$$

For RKBS $\mathcal{B}$:

- d_k: not expressible in terms of $k(x,y)$,

RKBS: computational intractability

Key for RKHS $\mathcal{H}_k$:

$$d_k{}^2(\mathbb{P},\mathbb{Q}) = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y).$$

For RKBS $\mathcal{B}$:

- d_k: not expressible in terms of $k(x,y)$,
- associated distances and estimators: no closed form expressions.

MMD: finished

Covariance operator

Idea: (un)centered cross-covariance

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[xy^{T}\right],$$

u: uncentered, c: centered.

Idea: (un)centered cross-covariance

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[xy^T\right], \qquad C_{xy}^{c} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

u: uncentered, c: centered.

Idea: (un)centered cross-covariance

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[xy^{T}\right], \qquad C_{xy}^{c} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^{T}\right],$$

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right],$$

u: uncentered, c: centered.

Idea: (un)centered cross-covariance

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[xy^T\right], \qquad C_{xy}^{c} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right], \quad C_{xy}^{c} = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

u: uncentered, c: centered.

Idea: (un)centered cross-covariance

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[xy^T\right], \qquad C_{xy}^{c} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$C_{xy}^{u} = \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right], \quad C_{xy}^{c} = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

u: uncentered, c: centered. In short, $xy^T \to \varphi(x) \otimes \psi(y)$.

Preview

$$C_{xy}^c = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

encodes the dependency of x and y.

- $C_{xy}^c = 0 \Leftrightarrow x \perp y$ for 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$.

Preview

$$C_{xy}^c = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

encodes the dependency of x and y.

- $C_{xy}^c = 0 \Leftrightarrow x \perp y$ for 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- $HSIC(x,y) = \left\|C_{xy}^c\right\|_{HS}$. $\|\cdot\|_{HS}$: extension of Frobenius norm.

Preview

$$C^c_{xy} = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

encodes the dependency of x and y.

- $C^c_{xy} = 0 \Leftrightarrow x \perp y$ for 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- $HSIC(x,y) = \left\|C^c_{xy}\right\|_{HS}$. $\|\cdot\|_{HS}$: extension of Frobenius norm.
- $HSIC(x,y)$: It will be easy to estimate. KCCA alternative.

Preview

$$C_{xy}^c = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

encodes the dependency of x and y.

- $C_{xy}^c = 0 \Leftrightarrow x \perp y$ for 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- $HSIC(x,y) = \left\|C_{xy}^c\right\|_{HS}$. $\|\cdot\|_{HS}$: extension of Frobenius norm.
- $HSIC(x,y)$: It will be easy to estimate. KCCA alternative.

Question

What is $\varphi(x) \otimes \psi(y)$ and $\|\cdot\|_{HS}$?

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right)$$

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

- Alternatively

$$\mathbb{R} \ni f^T \left(ab^T\right) g = \left(f^T a\right) \left(b^T g\right)$$

$ab^T : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ bilinear form.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a\left(b^T g\right) = a\langle b, g\rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

- Alternatively

$$\mathbb{R} \ni f^T\left(ab^T\right) g = \left(f^T a\right)\left(b^T g\right) = \langle f, a\rangle \langle g, b\rangle$$

$ab^T : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ bilinear form.

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}.$$

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2 \right\}.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}.$$

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2 \right\}.$$

- Define inner product on $\mathcal{L}$, extended by linearity

$$\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle := \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2}.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}.$$

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2 \right\}.$$

- Define inner product on $\mathcal{L}$, extended by linearity

$$\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle := \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2}.$$

- $\mathcal{H}_1 \otimes \mathcal{H}_2$: completion of $\mathcal{L}$.

$a_1 \otimes \ldots \otimes a_M$, $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_M$ would work similarly

Tensor product of M Hilbert spaces:

$$(a_1 \otimes \ldots \otimes a_M)(h_1, \ldots, h_M) = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m},$$

$a_1 \otimes \ldots \otimes a_M$, $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_M$ would work similarly

Tensor product of M Hilbert spaces:

$$(a_1 \otimes \ldots \otimes a_M)(h_1, \ldots, h_M) = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m},$$

$$\left\langle \otimes_{m=1}^{M} a_m, \otimes_{m=1}^{M} b_m \right\rangle = \prod_{m=1}^{M} \langle a_m, b_m \rangle_{\mathcal{H}_m}.$$

$\Rightarrow$ HSIC for M-variables.

$\langle \cdot, \cdot \rangle$: well-defined, pos. definite [Reed and Simon, 1980]

Well-defined: (λ, λ') is expansion-independent, i.e.

$$\lambda_1 = \sum_i c_i a_i \otimes b_i = \lambda_2 = \sum_j c'_j a'_j \otimes b'_j,$$

$$\langle \lambda_1, \lambda' \rangle \stackrel{?}{=} \langle \lambda_2, \lambda' \rangle$$

$\langle\cdot,\cdot\rangle$: well-defined, pos. definite [Reed and Simon, 1980]

Well-defined: (λ, λ') is expansion-independent, i.e.

$$\lambda_1 = \sum_i c_i a_i \otimes b_i = \lambda_2 = \sum_j c'_j a'_j \otimes b'_j,$$

$$\langle \lambda_1, \lambda' \rangle \stackrel{?}{=} \langle \lambda_2, \lambda' \rangle \Leftrightarrow \langle \underbrace{\lambda_1 - \lambda_2}_{=0}, \lambda' \rangle \stackrel{?}{=} 0 \quad (\forall \lambda' \in \mathcal{L}).$$

$\langle \cdot, \cdot \rangle$: well-defined, pos. definite [Reed and Simon, 1980]

Well-defined: (λ, λ') is expansion-independent, i.e.

$$\lambda_1 = \sum_i c_i a_i \otimes b_i = \lambda_2 = \sum_j c'_j a'_j \otimes b'_j,$$

$$\langle \lambda_1, \lambda' \rangle \stackrel{?}{=} \langle \lambda_2, \lambda' \rangle \Leftrightarrow \langle \underbrace{\lambda_1 - \lambda_2}_{=0}, \lambda' \rangle \stackrel{?}{=} 0 \quad (\forall \lambda' \in \mathcal{L}).$$

- In other words $n = 0 \stackrel{?}{\Rightarrow} \langle n, \lambda' \rangle = 0$, $\forall \lambda' \in \mathcal{L}$.

$\langle \cdot, \cdot \rangle$: well-defined, pos. definite [Reed and Simon, 1980]

Well-defined: (λ, λ') is expansion-independent, i.e.

$$\lambda_1 = \sum_i c_i a_i \otimes b_i = \lambda_2 = \sum_j c'_j a'_j \otimes b'_j,$$

$$\langle \lambda_1, \lambda' \rangle \stackrel{?}{=} \langle \lambda_2, \lambda' \rangle \Leftrightarrow \langle \underbrace{\lambda_1 - \lambda_2}_{=0}, \lambda' \rangle \stackrel{?}{=} 0 \quad (\forall \lambda' \in \mathcal{L}).$$

- In other words $n = 0 \stackrel{?}{\Rightarrow} \langle n, \lambda' \rangle = 0,\ \forall \lambda' \in \mathcal{L}$.
- $\lambda' := \sum_i d_i e_i \otimes f_i$,

$$\langle 0, \lambda' \rangle = \left\langle 0, \sum_i d_i e_i \otimes f_i \right\rangle = \sum_i d_i \underbrace{\langle 0, e_i \otimes f_i \rangle}_{=0(e_i, f_i)=0} = 0.$$

$\langle \cdot, \cdot \rangle$ is positive definite

- Goal: $\langle \lambda, \lambda \rangle = 0 \Rightarrow \lambda = 0$.

$\langle \cdot, \cdot \rangle$ is positive definite

- Goal: $\langle \lambda, \lambda \rangle = 0 \Rightarrow \lambda = 0$.
- $\lambda := \sum_i c_i a_i \otimes b_i$, $A := span\{(a_i)\} \subset \mathcal{H}_1$, $B := span\{(b_i)\} \subset \mathcal{H}_2$.

$\langle \cdot, \cdot \rangle$ is positive definite

- Goal: $\langle \lambda, \lambda \rangle = 0 \Rightarrow \lambda = 0$.
- $\lambda := \sum_i c_i a_i \otimes b_i$, $A := span\{(a_i)\} \subset \mathcal{H}_1$, $B := span\{(b_i)\} \subset \mathcal{H}_2$.
- (α_i) := ONB for A, (β_j) := ONB for B.

$\langle \cdot, \cdot \rangle$ is positive definite

- Goal: $\langle \lambda, \lambda \rangle = 0 \Rightarrow \lambda = 0$.
- $\lambda := \sum_i c_i a_i \otimes b_i$, $A := span\{(a_i)\} \subset \mathcal{H}_1$, $B := span\{(b_i)\} \subset \mathcal{H}_2$.
- $(\alpha_i) :=$ ONB for A, $(\beta_j) :=$ ONB for B.
- $a_i \in A$, $b_i \in B$ hence

$$\lambda = \sum_{i,j} c_{ij} \alpha_i \otimes \beta_j,$$

$$\langle \lambda, \lambda \rangle = \left\langle \sum_{i,j} c_{ij} \alpha_i \otimes \beta_j, \sum_{u,v} c_{uv} \alpha_u \otimes \beta_v \right\rangle$$

$\langle \cdot, \cdot \rangle$ is positive definite

- Goal: $\langle \lambda, \lambda \rangle = 0 \Rightarrow \lambda = 0$.
- $\lambda := \sum_i c_i a_i \otimes b_i$, $A := span\{(a_i)\} \subset \mathcal{H}_1$, $B := span\{(b_i)\} \subset \mathcal{H}_2$.
- $(\alpha_i) :=$ ONB for A, $(\beta_j) :=$ ONB for B.
- $a_i \in A$, $b_i \in B$ hence

$$\lambda = \sum_{i,j} c_{ij} \alpha_i \otimes \beta_j,$$

$$\begin{aligned} \langle \lambda, \lambda \rangle &= \left\langle \sum_{i,j} c_{ij} \alpha_i \otimes \beta_j, \sum_{u,v} c_{uv} \alpha_u \otimes \beta_v \right\rangle \\ &= \sum_{i,j,u,v} c_{ij} c_{uv} \underbrace{\langle \alpha_i \otimes \beta_j, \alpha_u, \otimes \beta_v \rangle}_{\langle \alpha_i, \alpha_u \rangle_{\mathcal{H}_1} \langle \beta_j, \beta_v \rangle_{\mathcal{H}_2} = \delta_{iu} \delta_{jv}} \end{aligned}$$

$\langle \cdot, \cdot \rangle$ is positive definite

- Goal: $\langle \lambda, \lambda \rangle = 0 \Rightarrow \lambda = 0$.
- $\lambda := \sum_i c_i a_i \otimes b_i$, $A := span\{(a_i)\} \subset \mathcal{H}_1$, $B := span\{(b_i)\} \subset \mathcal{H}_2$.
- $(\alpha_i) :=$ ONB for A, $(\beta_j) :=$ ONB for B.
- $a_i \in A$, $b_i \in B$ hence

$$\lambda = \sum_{i,j} c_{ij} \alpha_i \otimes \beta_j,$$

$$\langle \lambda, \lambda \rangle = \left\langle \sum_{i,j} c_{ij} \alpha_i \otimes \beta_j, \sum_{u,v} c_{uv} \alpha_u \otimes \beta_v \right\rangle$$

$$= \sum_{i,j,u,v} c_{ij} c_{uv} \underbrace{\langle \alpha_i \otimes \beta_j, \alpha_u, \otimes \beta_v \rangle}_{\langle \alpha_i, \alpha_u \rangle_{\mathcal{H}_1} \langle \beta_j, \beta_v \rangle_{\mathcal{H}_2} = \delta_{iu} \delta_{jv}} = \sum_{i,j} c_{i,j}^2.$$

$\langle\cdot,\cdot\rangle$ is positive definite

- Goal: $\langle\lambda,\lambda\rangle = 0 \Rightarrow \lambda = 0$.
- $\lambda := \sum_i c_i a_i \otimes b_i$, $A := span\{(a_i)\} \subset \mathcal{H}_1$, $B := span\{(b_i)\} \subset \mathcal{H}_2$.
- $(\alpha_i) :=$ ONB for A, $(\beta_j) :=$ ONB for B.
- $a_i \in A$, $b_i \in B$ hence

$$\lambda = \sum_{i,j} c_{ij}\alpha_i \otimes \beta_j,$$

$$\langle\lambda,\lambda\rangle = \left\langle \sum_{i,j} c_{ij}\alpha_i \otimes \beta_j, \sum_{u,v} c_{uv}\alpha_u \otimes \beta_v \right\rangle$$

$$= \sum_{i,j,u,v} c_{ij}c_{uv} \underbrace{\langle\alpha_i \otimes \beta_j, \alpha_u, \otimes\beta_v\rangle}_{\langle\alpha_i,\alpha_u\rangle_{\mathcal{H}_1}\langle\beta_j,\beta_v\rangle_{\mathcal{H}_2}=\delta_{iu}\delta_{jv}} = \sum_{i,j} c_{i,j}^2.$$

- In short, $\langle\lambda,\lambda\rangle = 0 \Rightarrow c_{ij} = 0$ $(\forall i,j)$, i.e. $\lambda = 0$.

Tensor product of RKHSs

Theorem ([Berlinet and Thomas-Agnan, 2004])

- *Given:* $\mathcal{H}_1 = \mathcal{H}_k$, $\mathcal{H}_2 = \mathcal{H}_\ell$ *RKHSs with kernel* k *and* ℓ.
- *Then* $\mathcal{H}_1 \otimes \mathcal{H}_2$ *is RKHS with kernel*

$$k \otimes \ell : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R},$$
$$(k \otimes \ell)\left((x_1, y_1), (x_2, y_2)\right) := k(x_1, x_2)\ell(y_1, y_2).$$

Tensor product of RKHSs

Theorem ([Berlinet and Thomas-Agnan, 2004])

- *Given:* $\mathcal{H}_1 = \mathcal{H}_k$, $\mathcal{H}_2 = \mathcal{H}_\ell$ *RKHSs with kernel* k *and* ℓ.
- *Then* $\mathcal{H}_1 \otimes \mathcal{H}_2$ *is RKHS with kernel*

$$k \otimes \ell : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R},$$
$$(k \otimes \ell)\left((x_1, y_1), (x_2, y_2)\right) := k(x_1, x_2)\ell(y_1, y_2).$$

Intuition:

- inner product on $\mathcal{X}$ and $\mathcal{Y}$ $\to$ inner product on $\mathcal{X} \times \mathcal{Y}$.
- $\mathcal{X}$ = animal images, $\mathcal{Y}$ = descriptions of animals.

Until now

- $a \otimes b$: defined; 'nice' operator (HS:=Hilbert-Schmidt).

Until now

- $a \otimes b$: defined; 'nice' operator (HS:=Hilbert-Schmidt).
- It will descend to its expectation ($C_{xy}^u \Rightarrow$ HSIC).

Until now

- $a \otimes b$: defined; 'nice' operator (HS:=Hilbert-Schmidt).
- It will descend to its expectation ($C_{xy}^u \Rightarrow$ HSIC).
- (e_i), (f_j): canonical basis in $\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$.
- HS operators: extensions of $L \in \mathbb{R}^{d_2 \times d_1}$ with

$$\|L\|_F^2 = \sum_i \| \underbrace{Le_i}_{i^{th} \text{ column of } L} \|_2^2$$

Until now

- $a \otimes b$: defined; 'nice' operator (HS:=Hilbert-Schmidt).
- It will descend to its expectation ($C_{xy}^u \Rightarrow$ HSIC).
- (e_i), (f_j): canonical basis in $\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$.
- HS operators: extensions of $L \in \mathbb{R}^{d_2 \times d_1}$ with

$$\|L\|_F^2 = \sum_i \|\underbrace{Le_i}_{i^{th} \text{ column of } L}\|_2^2 = \sum_i \sum_j (Le_i)_j^2$$

Until now

- $a \otimes b$: defined; 'nice' operator (HS:=Hilbert-Schmidt).
- It will descend to its expectation ($C_{xy}^u \Rightarrow$ HSIC).
- (e_i), (f_j): canonical basis in $\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$.
- HS operators: extensions of $L \in \mathbb{R}^{d_2 \times d_1}$ with

$$\|L\|_F^2 = \sum_i \| \underbrace{Le_i}_{i^{th} \text{ column of } L} \|_2^2 = \sum_i \sum_j (Le_i)_j^2 = \sum_{i,j} L_{ji}^2.$$

Hilbert-Schmidt operators: quick summary

- An $L : \mathcal{H}_1 \to \mathcal{H}_2$ bounded linear operator is called Hilbert-Schmidt if

$$\|L\|_{HS}^2 := \sum_i \underbrace{\|Le_i\|_{\mathcal{H}_2}^2}_{=\sum_j \left\langle Le_i, f_j \right\rangle_{\mathcal{H}_2}^2} < \infty.$$

Hilbert-Schmidt operators: quick summary

- $\mathcal{H}_1, \mathcal{H}_2$: separable Hilbert spaces. $(e_i)_{i\in I}$, $(f_j)_{j\in J}$: ONB in $\mathcal{H}_1$, $\mathcal{H}_2$.
- An $L : \mathcal{H}_1 \to \mathcal{H}_2$ bounded linear operator is called Hilbert-Schmidt if

$$\|L\|_{HS}^2 := \sum_i \underbrace{\|Le_i\|_{\mathcal{H}_2}^2}_{=\sum_j \langle Le_i, f_j \rangle_{\mathcal{H}_2}^2} < \infty.$$

Hilbert-Schmidt operators: quick summary

- $\mathcal{H}_1, \mathcal{H}_2$: separable Hilbert spaces. $(e_i)_{i\in I}$, $(f_j)_{j\in J}$: ONB in $\mathcal{H}_1$, $\mathcal{H}_2$.
- An $L : \mathcal{H}_1 \to \mathcal{H}_2$ bounded linear operator is called Hilbert-Schmidt if

$$\|L\|_{HS}^2 := \sum_i \underbrace{\|Le_i\|_{\mathcal{H}_2}^2}_{=\sum_j \langle Le_i, f_j\rangle_{\mathcal{H}_2}^2} < \infty.$$

- Inner product on $L_1, L_2 \in HS(\mathcal{H}_1, \mathcal{H}_2)$

$$\langle L_1, L_2\rangle_{HS} := \sum_i \langle L_1 e_i, L_2 e_i\rangle_{\mathcal{H}_2}.$$

Hilbert-Schmidt operators: quick summary

- $\mathcal{H}_1, \mathcal{H}_2$: separable Hilbert spaces. $(e_i)_{i\in I}$, $(f_j)_{j\in J}$: ONB in $\mathcal{H}_1$, $\mathcal{H}_2$.
- An $L : \mathcal{H}_1 \to \mathcal{H}_2$ bounded linear operator is called Hilbert-Schmidt if

$$\|L\|_{HS}^2 := \sum_i \underbrace{\|Le_i\|_{\mathcal{H}_2}^2}_{=\sum_j \langle Le_i, f_j\rangle_{\mathcal{H}_2}^2} < \infty.$$

- Inner product on $L_1, L_2 \in HS(\mathcal{H}_1, \mathcal{H}_2)$

$$\langle L_1, L_2\rangle_{HS} := \sum_i \langle L_1 e_i, L_2 e_i\rangle_{\mathcal{H}_2}.$$

- $HS(\mathcal{H}_1, \mathcal{H}_2)$: Hilbert space.

Hilbert-Schmidt operators: notes

- $\mathcal{H}_1, \mathcal{H}_2$: separable $\Rightarrow$ I, J: countable, i.e. 'sums'.

Hilbert-Schmidt operators: notes

- $\mathcal{H}_1, \mathcal{H}_2$: separable $\Rightarrow$ I, J: countable, i.e. 'sums'.
- $\langle L_1, L_2 \rangle_{HS}$: well-defined (independent of the chosen basis).

Hilbert-Schmidt operators: notes

- $\mathcal{H}_1, \mathcal{H}_2$: separable $\Rightarrow$ I, J: countable, i.e. 'sums'.
- $\langle L_1, L_2 \rangle_{HS}$: well-defined (independent of the chosen basis).
- For RKHSs ($\mathcal{H}_i$): $\mathcal{X}$: separable, k: continuous $\Rightarrow$ $\mathcal{H}_k$: separable [Steinwart and Christmann, 2008].

$a \otimes b$ is Hilbert-Schmidt: linear & bounded

For $a \otimes b$ with $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$:

- linearity: ✓
- boundedness ($c \in \mathcal{H}_2$):

$$\|(a \otimes b)c\|_{\mathcal{H}_1} = \left\|a \langle b, c\rangle_{\mathcal{H}_2}\right\|_{\mathcal{H}_1}$$

$a \otimes b$ is Hilbert-Schmidt: linear & bounded

For $a \otimes b$ with $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$:

- linearity: ✓
- boundedness ($c \in \mathcal{H}_2$):

$$\|(a \otimes b)c\|_{\mathcal{H}_1} = \left\|a \langle b, c\rangle_{\mathcal{H}_2}\right\|_{\mathcal{H}_1} = \left|\langle b, c\rangle_{\mathcal{H}_2}\right| \|a\|_{\mathcal{H}_1}$$

$a \otimes b$ is Hilbert-Schmidt: linear & bounded

For $a \otimes b$ with $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$:

- linearity: ✓
- boundedness ($c \in \mathcal{H}_2$):

$$\|(a \otimes b)c\|_{\mathcal{H}_1} = \left\|a \langle b, c\rangle_{\mathcal{H}_2}\right\|_{\mathcal{H}_1} = \left|\langle b, c\rangle_{\mathcal{H}_2}\right| \|a\|_{\mathcal{H}_1} \overset{\text{CBS}}{\leqslant} \|b\|_{\mathcal{H}_2}\|c\|_{\mathcal{H}_2}\|a\|_{\mathcal{H}_1}.$$

$a \otimes b$ is Hilbert-Schmidt: linear & bounded

For $a \otimes b$ with $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$:

- linearity: ✓
- boundedness ($c \in \mathcal{H}_2$):

$$\|(a \otimes b)c\|_{\mathcal{H}_1} = \left\|a \langle b, c\rangle_{\mathcal{H}_2}\right\|_{\mathcal{H}_1} = \left|\langle b, c\rangle_{\mathcal{H}_2}\right| \|a\|_{\mathcal{H}_1} \overset{\text{CBS}}{\leqslant} \|b\|_{\mathcal{H}_2}\|c\|_{\mathcal{H}_2}\|a\|_{\mathcal{H}_1}.$$

Thus $\|a \otimes b\| \leqslant \|a\|_{\mathcal{H}_1} \|b\|_{\mathcal{H}_2} < \infty$.

$a \otimes b$ is a Hilbert-Schmidt operator

Let $(e_i)_{i \in I} \subset \mathcal{H}_2$ ONB,

$$\|a \otimes b\|_{HS}^2 = \sum_i \left\|(a \otimes b)e_i\right\|_{\mathcal{H}_1}^2$$

$a \otimes b$ is a Hilbert-Schmidt operator

Let $(e_i)_{i \in I} \subset \mathcal{H}_2$ ONB,

$$\|a \otimes b\|_{HS}^2 = \sum_i \left\|(a \otimes b)e_i\right\|_{\mathcal{H}_1}^2 = \sum_i \left\|a \langle b, e_i\rangle_{\mathcal{H}_2}\right\|_{\mathcal{H}_1}^2$$

$a \otimes b$ is a Hilbert-Schmidt operator

Let $(e_i)_{i \in I} \subset \mathcal{H}_2$ ONB,

$$\|a \otimes b\|_{HS}^2 = \sum_i \left\|(a \otimes b)e_i\right\|_{\mathcal{H}_1}^2 = \sum_i \left\|a \langle b, e_i \rangle_{\mathcal{H}_2}\right\|_{\mathcal{H}_1}^2$$

$$= \|a\|_{\mathcal{H}_1}^2 \underbrace{\sum_i \left|\langle b, e_i \rangle_{\mathcal{H}_2}\right|^2}_{\|b\|_{\mathcal{H}_2}^2 \text{(Parseval equality)}} < \infty.$$

Uncentered cross-covariance operator

$$C_{xy}^{u} := \mathbb{E}_{xy}\big[\underbrace{\varphi(x) \otimes \psi(y)}_{\in HS(\mathcal{H}_\ell,\mathcal{H}_k)}\big] \in HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right).$$

Uncentered cross-covariance operator

$$C_{xy}^u := \mathbb{E}_{xy}\Big[\underbrace{\varphi(x) \otimes \psi(y)}_{\in HS(\mathcal{H}_\ell,\mathcal{H}_k)}\Big] \in HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right).$$

- 'Same' construction as $\mu_{\mathbb{P}}$: we changed $\mathcal{H}_k$ to $HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right)$.

Uncentered cross-covariance operator

$$C_{xy}^{u} := \mathbb{E}_{xy}\Big[\underbrace{\varphi(x)\otimes\psi(y)}_{\in HS(\mathcal{H}_\ell,\mathcal{H}_k)}\Big] \in HS\left(\mathcal{H}_\ell,\mathcal{H}_k\right).$$

- 'Same' construction as $\mu_{\mathbb{P}}$: we changed $\mathcal{H}_k$ to $HS\left(\mathcal{H}_\ell,\mathcal{H}_k\right)$.
- Bochner integral: $\exists C_{xy}^{u} \Leftrightarrow \mathbb{E}_{xy}\left\|\varphi(x)\otimes\psi(y)\right\|_{HS} < \infty$.

Uncentered cross-covariance operator

$$C_{xy}^{u} := \mathbb{E}_{xy}\Big[\underbrace{\varphi(x) \otimes \psi(y)}_{\in HS(\mathcal{H}_\ell,\mathcal{H}_k)}\Big] \in HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right).$$

- 'Same' construction as $\mu_{\mathbb{P}}$: we changed $\mathcal{H}_k$ to $HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right)$.
- Bochner integral: $\exists C_{xy}^{u} \Leftrightarrow \mathbb{E}_{xy}\|\varphi(x) \otimes \psi(y)\|_{HS} < \infty$.
- $\|\varphi(x) \otimes \psi(y)\|_{HS} = \|\varphi(x)\|_{\mathcal{H}_k} \|\psi(y)\|_{\mathcal{H}_\ell}$

Uncentered cross-covariance operator

$$C_{xy}^u := \mathbb{E}_{xy}\Big[\underbrace{\varphi(x) \otimes \psi(y)}_{\in HS(\mathcal{H}_\ell, \mathcal{H}_k)}\Big] \in HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right).$$

- 'Same' construction as $\mu_{\mathbb{P}}$: we changed $\mathcal{H}_k$ to $HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right)$.
- Bochner integral: $\exists C_{xy}^u \Leftrightarrow \mathbb{E}_{xy}\|\varphi(x) \otimes \psi(y)\|_{HS} < \infty$.
- $\|\varphi(x) \otimes \psi(y)\|_{HS} = \|\varphi(x)\|_{\mathcal{H}_k} \|\psi(y)\|_{\mathcal{H}_\ell} = \sqrt{k(x,x)}\sqrt{\ell(y,y)}$.

Uncentered cross-covariance operator

$$C_{xy}^{u} := \mathbb{E}_{xy}\Big[\underbrace{\varphi(x) \otimes \psi(y)}_{\in HS(\mathcal{H}_\ell,\mathcal{H}_k)} \Big] \in HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right).$$

- 'Same' construction as $\mu_{\mathbb{P}}$: we changed $\mathcal{H}_k$ to $HS\left(\mathcal{H}_\ell, \mathcal{H}_k\right)$.
- Bochner integral: $\exists C_{xy}^{u} \Leftrightarrow \mathbb{E}_{xy}\left\|\varphi(x) \otimes \psi(y)\right\|_{HS} < \infty$.
- $\left\|\varphi(x) \otimes \psi(y)\right\|_{HS} = \left\|\varphi(x)\right\|_{\mathcal{H}_k} \left\|\psi(y)\right\|_{\mathcal{H}_\ell} = \sqrt{k(x,x)}\sqrt{\ell(y,y)}$.
- Sufficient condition: k and ℓ are bounded.

Centered covariance operator [Baker, 1973]

Let $\mu_x := \mu_{\mathbb{P}_x}$, $\mu_y := \mu_{\mathbb{P}_y}$.

$$C^c_{xy} = \mathbb{E}_{xy}\left[\left(\varphi(x) - \underbrace{\mathbb{E}_x\varphi(x)}_{\mu_x}\right) \otimes \left(\psi(y) - \underbrace{\mathbb{E}_y\psi(y)}_{\mu_y}\right)\right]$$

Centered covariance operator [Baker, 1973]

Let $\mu_x := \mu_{\mathbb{P}_x}$, $\mu_y := \mu_{\mathbb{P}_y}$.

$$C_{xy}^c = \mathbb{E}_{xy}\left[\left(\varphi(x) - \underbrace{\mathbb{E}_x\varphi(x)}_{\mu_x}\right) \otimes \left(\psi(y) - \underbrace{\mathbb{E}_y\psi(y)}_{\mu_y}\right)\right]$$

$$= \underbrace{\mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right]}_{C_{xy}^u \in HS(\mathcal{H}_\ell, \mathcal{H}_k)} - \underbrace{\mu_x \otimes \mu_y}_{\in HS(\mathcal{H}_\ell, \mathcal{H}_k)} \in HS(\mathcal{H}_\ell, \mathcal{H}_k).$$

Hilbert-Schmidt independence criterion (HSIC)

HSIC [Fukumizu et al., 2004, Gretton et al., 2005a]:

$$HSIC(x, y; \mathcal{H}_k, \mathcal{H}_\ell) := \left\| C_{xy}^c \right\|_{HS}.$$

Hilbert-Schmidt independence criterion (HSIC)

HSIC [Fukumizu et al., 2004, Gretton et al., 2005a]:

$$HSIC(x, y; \mathcal{H}_k, \mathcal{H}_\ell) := \left\| C^c_{xy} \right\|_{HS}.$$

Question

When does HSIC characterize independence?

Hilbert-Schmidt independence criterion (HSIC)

HSIC [Fukumizu et al., 2004, Gretton et al., 2005a]:

$$HSIC(x, y; \mathcal{H}_k, \mathcal{H}_\ell) := \left\| C_{xy}^c \right\|_{HS}.$$

Question

When does HSIC characterize independence?
We will discuss it later (after HSIC $\Leftrightarrow$ distance covariance).

How do covariance operators encode covariance?

Let $g \in \mathcal{H}_\ell$, $f \in \mathcal{H}_k$, $HS := HS(\mathcal{H}_\ell, \mathcal{H}_k)$.

$$\left\langle f, C_{xy}^u g \right\rangle_{\mathcal{H}_k} = \left\langle C_{xy}^u, f \otimes g \right\rangle_{HS}$$

Cheating:

- next slide.
- Enough $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS\left(\mathcal{H}_2, \mathcal{H}_1\right)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)}$$

How do covariance operators encode covariance?

Let $g \in \mathcal{H}_\ell$, $f \in \mathcal{H}_k$, $HS := HS(\mathcal{H}_\ell, \mathcal{H}_k)$.

$$\left\langle f, C_{xy}^u g \right\rangle_{\mathcal{H}_k} = \left\langle C_{xy}^u, f \otimes g \right\rangle_{HS} = \langle \mathbb{E}_{xy}[\varphi(x) \otimes \psi(y)], f \otimes g \rangle_{HS}$$

Cheating:

- next slide.
- Enough $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS\left(\mathcal{H}_2, \mathcal{H}_1\right)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2, \mathcal{H}_1)}$$

How do covariance operators encode covariance?

Let $g \in \mathcal{H}_\ell$, $f \in \mathcal{H}_k$, $HS := HS(\mathcal{H}_\ell, \mathcal{H}_k)$.

$$\left\langle f, C^u_{xy} g \right\rangle_{\mathcal{H}_k} = \left\langle C^u_{xy}, f \otimes g \right\rangle_{HS} = \langle \mathbb{E}_{xy}[\varphi(x) \otimes \psi(y)], f \otimes g \rangle_{HS}$$
$$= \mathbb{E}_{xy} \underbrace{\langle \varphi(x) \otimes \psi(y), f \otimes g \rangle_{HS}}_{= f(x) g(y)}$$

Cheating:

- next slide.
- Enough $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2, \mathcal{H}_1)}$$

How do covariance operators encode covariance?

Let $g \in \mathcal{H}_\ell$, $f \in \mathcal{H}_k$, $HS := HS(\mathcal{H}_\ell, \mathcal{H}_k)$.

$$\left\langle f, C_{xy}^u g \right\rangle_{\mathcal{H}_k} = \left\langle C_{xy}^u, f \otimes g \right\rangle_{HS} = \langle \mathbb{E}_{xy}[\varphi(x) \otimes \psi(y)], f \otimes g \rangle_{HS}$$
$$= \mathbb{E}_{xy} \underbrace{\langle \varphi(x) \otimes \psi(y), f \otimes g \rangle_{HS}}_{=f(x)g(y)} = \mathbb{E}_{xy} \left[f(x) g(y) \right].$$

Cheating:

- next slide.
- Enough $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2, \mathcal{H}_1)}$$

Cheating

Statement: with $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)}.$$

Proof: $(b_i)_{i \in I}$ ONB in $\mathcal{H}_2$,

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \left\langle f, L \sum_i \langle g, b_i \rangle_{\mathcal{H}_2} b_i \right\rangle_{\mathcal{H}_1}$$

Cheating

Statement: with $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)}.$$

Proof: $(b_i)_{i \in I}$ ONB in $\mathcal{H}_2$,

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \left\langle f, L \sum_i \langle g, b_i \rangle_{\mathcal{H}_2} b_i \right\rangle_{\mathcal{H}_1} = \sum_i \langle g, b_i \rangle_{\mathcal{H}_2} \langle f, Lb_i \rangle_{\mathcal{H}_1}$$

Cheating

Statement: with $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)}.$$

Proof: $(b_i)_{i \in I}$ ONB in $\mathcal{H}_2$,

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \left\langle f, L\sum_i \langle g, b_i \rangle_{\mathcal{H}_2} b_i \right\rangle_{\mathcal{H}_1} = \sum_i \langle g, b_i \rangle_{\mathcal{H}_2} \langle f, Lb_i \rangle_{\mathcal{H}_1}$$

$$\stackrel{\otimes}{=} \sum_i \langle Lb_i, (f \otimes g) b_i \rangle_{\mathcal{H}_1}$$

Cheating

Statement: with $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2, \mathcal{H}_1)}.$$

Proof: $(b_i)_{i \in I}$ ONB in $\mathcal{H}_2$,

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \left\langle f, L \sum_i \langle g, b_i \rangle_{\mathcal{H}_2} b_i \right\rangle_{\mathcal{H}_1} = \sum_i \langle g, b_i \rangle_{\mathcal{H}_2} \langle f, Lb_i \rangle_{\mathcal{H}_1}$$

$$\overset{\otimes}{=} \sum_i \langle Lb_i, (f \otimes g) b_i \rangle_{\mathcal{H}_1} \overset{\langle \cdot, \cdot \rangle_{HS}}{=} \langle L, f \otimes g \rangle_{HS}.$$

Cheating

Statement: with $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)}.$$

With $L := a \otimes b$

$$\langle a \otimes b, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)} = \langle f, (a \otimes b)g \rangle_{\mathcal{H}_1}$$

Cheating

Statement: with $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$, $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$

$$\langle f, Lg \rangle_{\mathcal{H}_1} = \langle L, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)}.$$

With $L := a \otimes b$

$$\langle a \otimes b, f \otimes g \rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)} = \langle f, (a \otimes b)g \rangle_{\mathcal{H}_1} \stackrel{\otimes}{=} \langle a, f \rangle_{\mathcal{H}_1} \langle b, g \rangle_{\mathcal{H}_2}.$$

Remember: we have seen this for $a = f$, $b = g$.

Effect of the centered cross-covariance operator

Using that $C_{xy}^c = C_{xy}^u - \mu_x \otimes \mu_y$

$$\left\langle f, C_{xy}^c g \right\rangle_{\mathcal{H}_k} = \left\langle f, C_{xy}^u g \right\rangle_{\mathcal{H}_k} - \langle f, (\mu_x \otimes \mu_y) g \rangle_{\mathcal{H}_k}$$

Effect of the centered cross-covariance operator

Using that $C_{xy}^c = C_{xy}^u - \mu_x \otimes \mu_y$

$$\left\langle f, C_{xy}^c g \right\rangle_{\mathcal{H}_k} = \left\langle f, C_{xy}^u g \right\rangle_{\mathcal{H}_k} - \langle f, (\mu_x \otimes \mu_y) g \rangle_{\mathcal{H}_k}$$

$$\stackrel{\otimes}{=} \mathbb{E}_{xy}[f(x)g(y)] - \underbrace{\langle f, \mu_x \rangle_{\mathcal{H}_k}}_{\mathbb{E}_x f(x)} \underbrace{\langle g, \mu_y \rangle_{\mathcal{H}_\ell}}_{\mathbb{E}_y g(y)}$$

Effect of the centered cross-covariance operator

Using that $C_{xy}^c = C_{xy}^u - \mu_x \otimes \mu_y$

$$\left\langle f, C_{xy}^c g \right\rangle_{\mathcal{H}_k} = \left\langle f, C_{xy}^u g \right\rangle_{\mathcal{H}_k} - \left\langle f, (\mu_x \otimes \mu_y) g \right\rangle_{\mathcal{H}_k}$$

$$\stackrel{\otimes}{=} \mathbb{E}_{xy}[f(x)g(y)] - \underbrace{\langle f, \mu_x \rangle_{\mathcal{H}_k}}_{\mathbb{E}_x f(x)} \underbrace{\langle g, \mu_y \rangle_{\mathcal{H}_\ell}}_{\mathbb{E}_y g(y)}$$

$$= cov(f(x), g(y)).$$

Three notes

- KCCA formulation: using C^c_{xy}, C^c_{xx}, C^c_{yy}.

Three notes

- KCCA formulation: using C^c_{xy}, C^c_{xx}, C^c_{yy}.
- HSIC: captures $\mathbb{P}_{xy} \stackrel{?}{=} \mathbb{P}_x \otimes \mathbb{P}_y$ in $\mathcal{H}_k \otimes \mathcal{H}_\ell$.

Three notes

- KCCA formulation: using C^c_{xy}, C^c_{xx}, C^c_{yy}.
- HSIC: captures $\mathbb{P}_{xy} \stackrel{?}{=} \mathbb{P}_x \otimes \mathbb{P}_y$ in $\mathcal{H}_k \otimes \mathcal{H}_\ell$.
- Link to distance covariance, energy distance.

In other words, ...

KCCA formulation with cross-covariance operators

$$\rho_{\text{KCCA}}(x,y) = \sup_{f\in\mathcal{H}_k, g\in\mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)) \Leftrightarrow$$

$$\sup_{f\in\mathcal{H}_k, g\in\mathcal{H}_\ell} \left\langle f, C_{xy}^c g \right\rangle_{\mathcal{H}_k} \text{ s.t. } \begin{cases} \left\langle f, C_{xx}^c f \right\rangle_{\mathcal{H}_k} & = 1, \\ \left\langle g, C_{yy}^c g \right\rangle_{\mathcal{H}_\ell} & = 1. \end{cases}$$

KCCA: with κ-regularization

$$\rho_{\text{KCCA}}(x, y, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y); \kappa),$$

$$\operatorname{corr}(f(x), g(y); \kappa) = \frac{\operatorname{cov}_{xy}(f(x), g(y))}{\sqrt{\operatorname{var}_x f(x) + \kappa \|f\|^2_{\mathcal{H}_k}} \sqrt{\operatorname{var}_y g(y) + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

KCCA: with κ-regularization

$$\rho_{\text{KCCA}}(x, y, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y); \kappa),$$

$$\operatorname{corr}(f(x), g(y); \kappa) = \frac{\operatorname{cov}_{xy}(f(x), g(y))}{\sqrt{\operatorname{var}_x f(x) + \kappa \|f\|_{\mathcal{H}_k}^2} \sqrt{\operatorname{var}_y g(y) + \kappa \|g\|_{\mathcal{H}_\ell}^2}}.$$

Empirically,

$$\sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \left\langle f, \widehat{C_{xy}^c} g \right\rangle_{\mathcal{H}_k} \text{ s.t. } \begin{cases} \left\langle f, \left(\widehat{C_{xx}^c} + \kappa I\right) f \right\rangle_{\mathcal{H}_k} & = 1, \\ \left\langle g, \left(\widehat{C_{yy}^c} + \kappa I\right) g \right\rangle_{\mathcal{H}_\ell} & = 1. \end{cases}$$

KCCA: with κ-regularization

$$\rho_{\text{KCCA}}(x, y, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y); \kappa),$$

$$\operatorname{corr}(f(x), g(y); \kappa) = \frac{\operatorname{cov}_{xy}(f(x), g(y))}{\sqrt{\operatorname{var}_x f(x) + \kappa \|f\|^2_{\mathcal{H}_k}} \sqrt{\operatorname{var}_y g(y) + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

Empirically,

$$\sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \left\langle f, \widehat{C^c_{xy}} g \right\rangle_{\mathcal{H}_k} \text{ s.t. } \begin{cases} \left\langle f, \left(\widehat{C^c_{xx}} + \kappa I\right) f \right\rangle_{\mathcal{H}_k} & = 1, \\ \left\langle g, \left(\widehat{C^c_{yy}} + \kappa I\right) g \right\rangle_{\mathcal{H}_\ell} & = 1. \end{cases}$$

KCCA consistency analysis [Fukumizu et al., 2007]

using this formulation & the convergence of $\widehat{C^c_{xy}}$, $\widehat{C^c_{xx}}$, $\widehat{C^c_{yy}}$.

We saw $h((x,y),(x',y')) = k(x,x')\ell(y,y')$ is a kernel on $\mathcal{H}_k \otimes \mathcal{H}_\ell$. Let

$$\left\| \mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \otimes \mathbb{P}_y} \right\|_{\mathcal{H}_h}$$

HSIC: $\mathbb{P}_{xy} \stackrel{?}{=} \mathbb{P}_x \otimes \mathbb{P}_y$ in $\mathcal{H}_k \otimes \mathcal{H}_\ell$

We saw $h((x,y),(x',y')) = k(x,x')\ell(y,y')$ is a kernel on $\mathcal{H}_k \otimes \mathcal{H}_\ell$. Let

$$\left\|\mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \otimes \mathbb{P}_y}\right\|_{\mathcal{H}_h} = \Bigg\| \underbrace{\mathbb{E}_{\mathbb{P}_{xy}}\left[k\left(\cdot_1, x\right)\ell\left(\cdot_2, y\right)\right]}_{\mathbb{E}_{xy}[\varphi(x) \otimes \psi(y)]} - \underbrace{\mathbb{E}_{\mathbb{P}_x \mathbb{P}_y}\left[k\left(\cdot_1, x\right)\ell\left(\cdot_2, y\right)\right]}_{\mathbb{E}_x k(\cdot,x) \otimes \mathbb{E}_y \ell(\cdot,y)} \Bigg\|_{\mathcal{H}_h}$$

We saw $h((x,y),(x',y')) = k(x,x')\ell(y,y')$ is a kernel on $\mathcal{H}_k \otimes \mathcal{H}_\ell$. Let

$$\left\|\mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \otimes \mathbb{P}_y}\right\|_{\mathcal{H}_h} = \left\| \underbrace{\mathbb{E}_{\mathbb{P}_{xy}}\left[k\left(\cdot_1, x\right)\ell\left(\cdot_2, y\right)\right]}_{\mathbb{E}_{xy}[\varphi(x)\otimes\psi(y)]} - \underbrace{\mathbb{E}_{\mathbb{P}_x\mathbb{P}_y}\left[k\left(\cdot_1, x\right)\ell\left(\cdot_2, y\right)\right]}_{\mathbb{E}_x k(\cdot,x)\otimes\mathbb{E}_y \ell(\cdot,y)} \right\|_{\mathcal{H}_h}$$

$$= \left\|\mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right] - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_h}$$

HSIC: $\mathbb{P}_{xy} \stackrel{?}{=} \mathbb{P}_x \otimes \mathbb{P}_y$ in $\mathcal{H}_k \otimes \mathcal{H}_\ell$

We saw $h((x,y),(x',y')) = k(x,x')\ell(y,y')$ is a kernel on $\mathcal{H}_k \otimes \mathcal{H}_\ell$. Let

$$\left\|\mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x\otimes\mathbb{P}_y}\right\|_{\mathcal{H}_h} = \Big\| \underbrace{\mathbb{E}_{\mathbb{P}_{xy}}\left[k\left(\cdot_1,x\right)\ell\left(\cdot_2,y\right)\right]}_{\mathbb{E}_{xy}[\varphi(x)\otimes\psi(y)]} - \underbrace{\mathbb{E}_{\mathbb{P}_x\mathbb{P}_y}\left[k\left(\cdot_1,x\right)\ell\left(\cdot_2,y\right)\right]}_{\mathbb{E}_x k(\cdot,x)\otimes\mathbb{E}_y \ell(\cdot,y)} \Big\|_{\mathcal{H}_h}$$

$$= \left\|\mathbb{E}_{xy}\left[\varphi(x)\otimes\psi(y)\right] - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_h} = HSIC(x,y)$$

using $\mathcal{H}_1 \otimes \mathcal{H}_2 \simeq HS(\mathcal{H}_2,\mathcal{H}_1)$.

Distance covariance

- Characteristic functions: ϕ_{xy}, ϕ_x, ϕ_y.

Distance covariance

- Characteristic functions: ϕ_{xy}, ϕ_x, ϕ_y.
- Idea [Székely et al., 2007, Székely and Rizzo, 2009]:

$$x \perp y \Leftrightarrow \phi_{xy} = \phi_x \phi_y, \quad (x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}).$$

Distance covariance

- Characteristic functions: ϕ_{xy}, ϕ_x, ϕ_y.
- Idea [Székely et al., 2007, Székely and Rizzo, 2009]:

$$x \perp y \Leftrightarrow \phi_{xy} = \phi_x \phi_y, \quad (x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}).$$

- L^2_w norm of ϕ_{xy} and $\phi_x \phi_y$:

$$dCov(x, y) = \|\phi_{xy} - \phi_x \phi_y\|_{L^2_w}$$

Distance covariance

- Characteristic functions: ϕ_{xy}, ϕ_x, ϕ_y.
- Idea [Székely et al., 2007, Székely and Rizzo, 2009]:

$$x \perp y \Leftrightarrow \phi_{xy} = \phi_x \phi_y, \quad (x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}).$$

- L^2_w norm of ϕ_{xy} and $\phi_x \phi_y$, $\alpha \in (0,2)$:

$$dCov(x,y) = \left\| \phi_{xy} - \phi_x \phi_y \right\|_{L^2_w}$$

$$w(a,b) = \frac{1}{c(d_1,\alpha)c(d_2,\alpha) \left\| a \right\|_2^{d_1+\alpha} \left\| b \right\|_2^{d_2+\alpha}},$$

Distance covariance

- Characteristic functions: ϕ_{xy}, ϕ_x, ϕ_y.
- Idea [Székely et al., 2007, Székely and Rizzo, 2009]:

$$x \perp y \Leftrightarrow \phi_{xy} = \phi_x \phi_y, \quad (x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}).$$

- L_w^2 norm of ϕ_{xy} and $\phi_x \phi_y$, $\alpha \in (0,2)$:

$$dCov(x,y) = \left\| \phi_{xy} - \phi_x \phi_y \right\|_{L_w^2}$$

$$w(a,b) = \frac{1}{c(d_1,\alpha) c(d_2,\alpha) \left\| a \right\|_2^{d_1+\alpha} \left\| b \right\|_2^{d_2+\alpha}},$$

$$c(d,\alpha) = \frac{2\pi^{\frac{d}{2}} \Gamma\left(1 - \frac{\alpha}{2}\right)}{\alpha 2^{\alpha} \Gamma\left(\frac{d+\alpha}{2}\right)}.$$

Distance covariance

- Characteristic functions: ϕ_{xy}, ϕ_x, ϕ_y.
- Idea [Székely et al., 2007, Székely and Rizzo, 2009]:

$$x \perp y \Leftrightarrow \phi_{xy} = \phi_x \phi_y, \quad (x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}).$$

- L_w^2 norm of ϕ_{xy} and $\phi_x\phi_y$, $\alpha \in (0,2)$:

$$dCov(x,y) = \left\| \phi_{xy} - \phi_x\phi_y \right\|_{L_w^2}$$

$$w(a,b) = \frac{1}{c(d_1,\alpha)c(d_2,\alpha)\left\|a\right\|_2^{d_1+\alpha}\left\|b\right\|_2^{d_2+\alpha}},$$

$$c(d,\alpha) = \frac{2\pi^{\frac{d}{2}}\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}.$$

- $x \perp y$ iff. $dCov(x,y) = 0$.

Distance covariance: $\alpha = 1$

Alternative form in terms of pairwise distances:

$$dCov^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\|x-x'\|_2\|y-y'\|_2 + \mathbb{E}_{xx'}\|x-x'\|_2\mathbb{E}_{yy'}\|y-y'\|_2 - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\|x-x'\|_2\mathbb{E}_{y'}\|y-y'\|_2\right].$$

Distance covariance: $\alpha = 1$

Alternative form in terms of pairwise distances:

$$dCov^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\left\|x - x'\right\|_2\left\|y - y'\right\|_2 + \mathbb{E}_{xx'}\left\|x - x'\right\|_2\mathbb{E}_{yy'}\left\|y - y'\right\|_2 - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\left\|x - x'\right\|_2\mathbb{E}_{y'}\left\|y - y'\right\|_2\right].$$

Extension [Lyons, 2013]:

$$dCov^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\rho_1\left(x, x'\right)\rho_2\left(y, y'\right) + \mathbb{E}_{xx'}\left(x, x'\right)\mathbb{E}_{yy'}\left(y, y'\right) - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\rho_1\left(x, x'\right)\mathbb{E}_{y'}\rho_2\left(y, y'\right)\right],$$

Distance covariance: $\alpha = 1$

Alternative form in terms of pairwise distances:

$$dCov^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\left\|x - x'\right\|_2\left\|y - y'\right\|_2 + \mathbb{E}_{xx'}\left\|x - x'\right\|_2\mathbb{E}_{yy'}\left\|y - y'\right\|_2 - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\left\|x - x'\right\|_2\mathbb{E}_{y'}\left\|y - y'\right\|_2\right].$$

Extension [Lyons, 2013]:

$$dCov^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\rho_1\left(x, x'\right)\rho_2\left(y, y'\right) + \mathbb{E}_{xx'}\left(x, x'\right)\mathbb{E}_{yy'}\left(y, y'\right) - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\rho_1\left(x, x'\right)\mathbb{E}_{y'}\rho_2\left(y, y'\right)\right],$$

$(\mathcal{X}, \rho_1)$, $(\mathcal{Y}, \rho_2)$: metric spaces of negative type.

Distance covariance vs. HSIC

$$dCov^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\rho_1\left(x,x'\right)\rho_2\left(y,y'\right) + \mathbb{E}_{xx'}\rho_1\left(x,x'\right)\mathbb{E}_{yy'}\rho_2\left(y,y'\right) \\ - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\rho_1\left(x,x'\right)\mathbb{E}_{y'}\rho_2\left(y,y'\right)\right].$$

Distance covariance vs. HSIC

$$dCov^2(x, y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}\rho_1\left(x, x'\right)\rho_2\left(y, y'\right) + \mathbb{E}_{xx'}\rho_1\left(x, x'\right)\mathbb{E}_{yy'}\rho_2\left(y, y'\right) - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}\rho_1\left(x, x'\right)\mathbb{E}_{y'}\rho_2\left(y, y'\right)\right].$$

Recall:

$$HSIC^2(x, y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x, x')\ell(y, y') + \mathbb{E}_{xx'}k(x, x')\mathbb{E}_{yy'}\ell(y, y') - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x, x')\mathbb{E}_{y'}\ell(y, y')\right].$$

HSIC $\Leftrightarrow$ distance covariance

+extension to semi-metric spaces of negative type:

Theorem ([Sejdinovic et al., 2013b])

$dCov^2(x, y; \rho_1, \rho_2) = 4HSIC^2(x, y; \mathcal{H}_k, \mathcal{H}_\ell)$, *where*

$$\rho_1(x, x') = k(x, x) + k(x', x') - 2k(x, x'),$$
$$\rho_2(y, y') = \ell(y, y) + \ell(y', y') - 2\ell(y, y').$$

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.
- Triangle inequality: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$ for $\forall x, y, z \in \mathcal{X}$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.
- Triangle inequality: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$ for $\forall x, y, z \in \mathcal{X}$.

Examples:

- $\mathcal{X} = \mathbb{R}^d$, $\rho(x, y) = \|x - y\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p\right)^{\frac{1}{p}}$, $p \geqslant 1$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.
- Triangle inequality: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$ for $\forall x, y, z \in \mathcal{X}$.

Examples:

- $\mathcal{X} = \mathbb{R}^d$, $\rho(x, y) = \|x - y\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p\right)^{\frac{1}{p}}$, $p \geqslant 1$.
- $\mathcal{X} = C[a, b]$, $\rho(x, y) = \max_{z \in [a,b]} |x(z) - y(z)|$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.
- Triangle inequality: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$ for $\forall x, y, z \in \mathcal{X}$.

Examples:

- $\mathcal{X} = \mathbb{R}^d$, $\rho(x, y) = \|x - y\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p\right)^{\frac{1}{p}}$, $p \geqslant 1$.
- $\mathcal{X} = C[a, b]$, $\rho(x, y) = \max_{z \in [a,b]} |x(z) - y(z)|$.
- $\mathcal{X}$ any set. $\rho(x, y) = \delta_{x=y}$.

Semi-metric space: no triangle inequality

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a semi-metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- symmetry: $\rho(x, y) = \rho(y, x)$, for $\forall x, y \in \mathcal{X}$.

Semi-metric space: no triangle inequality

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a semi-metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- symmetry: $\rho(x, y) = \rho(y, x)$, for $\forall x, y \in \mathcal{X}$.

It is called negative type if in addition

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \rho(x_i, x_j) \leqslant 0$$

for $\forall n \geqslant 2$, $\forall x_1, \ldots, x_n \in \mathcal{X}$ and $\forall a_1, \ldots, a_n \in \mathbb{R}$ with $\sum_{i=1}^{n} a_i = 0$.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0, 1)$.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$
- $+$1st part $\Rightarrow \rho\,(x,y) = \|x - y\|_2^q \checkmark$ with $q \in (0,2)$.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$
- +1st part $\Rightarrow \rho\,(x,y) = \|x-y\|_2^q \checkmark$ with $q \in (0,2)$.
- Specifically: $\rho\,(x,y) = \|x-y\|_2$ is OK.

Energy distance [Székely and Rizzo, 2004, Baringhaus and Franz, 2004, Székely and Rizzo, 2005]

$x, x' \sim \mathbb{P}$, $y, y' \sim \mathbb{Q}$:

$$EnDist(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{xy} \|x - y\|_2 - \mathbb{E}_{xx'} \left\|x - x'\right\|_2 - \mathbb{E}_{yy'} \left\|y - y'\right\|_2,$$

Energy distance [Székely and Rizzo, 2004, Baringhaus and Franz, 2004, Székely and Rizzo, 2005]

$x, x' \sim \mathbb{P}$, $y, y' \sim \mathbb{Q}$:

$$EnDist(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{xy}\|x - y\|_2 - \mathbb{E}_{xx'}\left\|x - x'\right\|_2 - \mathbb{E}_{yy'}\left\|y - y'\right\|_2,$$

$$EnDist(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{xy}\rho\left(x, y\right) - \mathbb{E}_{xx'}\rho\left(x, x'\right) - \mathbb{E}_{yy'}\rho\left(y, y'\right).$$

Energy distance [Székely and Rizzo, 2004, Baringhaus and Franz, 2004, Székely and Rizzo, 2005]

$x, x' \sim \mathbb{P}$, $y, y' \sim \mathbb{Q}$:

$$EnDist(\mathbb{P},\mathbb{Q}) = 2\mathbb{E}_{xy}\|x-y\|_2 - \mathbb{E}_{xx'}\left\|x-x'\right\|_2 - \mathbb{E}_{yy'}\left\|y-y'\right\|_2,$$
$$EnDist(\mathbb{P},\mathbb{Q}) = 2\mathbb{E}_{xy}\rho\left(x,y\right) - \mathbb{E}_{xx'}\rho\left(x,x'\right) - \mathbb{E}_{yy'}\rho\left(y,y'\right).$$

Properties:

- $EnDist(\mathbb{P},\mathbb{Q}) \geqslant 0$ with ρ metric of negative-type.

Energy distance [Székely and Rizzo, 2004, Baringhaus and Franz, 2004, Székely and Rizzo, 2005]

$x, x' \sim \mathbb{P}$, $y, y' \sim \mathbb{Q}$:

$$EnDist(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{xy}\|x - y\|_2 - \mathbb{E}_{xx'}\left\|x - x'\right\|_2 - \mathbb{E}_{yy'}\left\|y - y'\right\|_2,$$
$$EnDist(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{xy}\rho(x, y) - \mathbb{E}_{xx'}\rho\left(x, x'\right) - \mathbb{E}_{yy'}\rho\left(y, y'\right).$$

Properties:

- $EnDist(\mathbb{P}, \mathbb{Q}) \geqslant 0$ with ρ metric of negative-type.
- $EnDist(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ for $(\mathcal{X}, \rho)$ strictly negative spaces; example: $(\mathbb{R}^d, \|\cdot\|_2)$.

Strict negativity

In addition:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \rho(x_i, x_j) < 0$$

if x_i-s are distinct and $\exists a_i \neq 0$.

Energy distance vs. MMD

Energy distance: also called N-distance [Zinger et al., 1992, Klebanov, 2005],

$$EnDist(\mathbb{P},\mathbb{Q}) = 2\mathbb{E}_{xy}\rho\left(x,y\right) - \mathbb{E}_{xx'}\rho\left(x,x'\right) - \mathbb{E}_{yy'}\rho\left(y,y'\right).$$

Energy distance vs. MMD

Energy distance: also called N-distance
[Zinger et al., 1992, Klebanov, 2005],

$$EnDist(\mathbb{P},\mathbb{Q}) = 2\mathbb{E}_{xy}\rho\left(x,y\right) - \mathbb{E}_{xx'}\rho\left(x,x'\right) - \mathbb{E}_{yy'}\rho\left(y,y'\right).$$

MMD (recall):

$$MMD^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{x,x'}k(x,x') + \mathbb{E}_{y,y'}k(y,y') - 2\mathbb{E}_{xy}k(x,y).$$

MMD ⇔ energy distance

Theorem ([Sejdinovic et al., 2013b])

$$EnDist(\mathbb{P}, \mathbb{Q}; \rho) = 2MMD^2(\mathbb{P}, \mathbb{Q}; \mathcal{H}_k),$$

where

$$\rho(x, y) = k(x, x) + k(y, y) - 2k(x, y).$$

Central in applications: characteristic property

- HSIC, $k = \otimes_{m=1}^{M} k_m$, $x = (x_m)_{m=1}^{M}$:

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right), \quad k(x, x') := \prod_{m=1}^{M} k_m(x_m, x'_m).$$

Central in applications: characteristic property

- HSIC, $k = \otimes_{m=1}^{M} k_m$, $x = (x_m)_{m=1}^{M}$:

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M}\mathbb{P}_m\right), \quad k(x, x') := \prod_{m=1}^{M} k_m(x_m, x'_m).$$

$k = \otimes_{m=1}^{M} k_m$ will be called $\mathcal{I}$-characteristic if

$$\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^{M}\mathbb{P}_m.$$

Central in applications: characteristic property

- HSIC, $k = \otimes_{m=1}^{M} k_m$, $x = (x_m)_{m=1}^{M}$:

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M}\mathbb{P}_m\right), \quad k(x, x') := \prod_{m=1}^{M} k_m(x_m, x'_m).$$

$k = \otimes_{m=1}^{M} k_m$ will be called $\mathcal{I}$-characteristic if

$$\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^{M}\mathbb{P}_m.$$

Recall (MMD): k is called characteristic if

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Central in applications: characteristic property

- HSIC, $k = \otimes_{m=1}^{M} k_m$, $x = (x_m)_{m=1}^{M}$:

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M}\mathbb{P}_m\right), \quad k(x, x') := \prod_{m=1}^{M} k_m(x_m, x'_m).$$

$k = \otimes_{m=1}^{M} k_m$ will be called $\mathcal{I}$-characteristic if

$$\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^{M}\mathbb{P}_m.$$

Recall (MMD): k is called characteristic if

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

$\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic.
Relation? Conditions in terms of k_m-s?

$\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-char $\Longleftarrow$ char $\Longleftarrow$ universal

$(k_m)_{m=1}^{M}$: char $\overset{\text{[Sriperumbudur et al., 2011]}}{\not\Longrightarrow}$ / $\underset{\text{[Sriperumbudur et al., 2011]}}{\Longleftarrow}$ -universal

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:
  $k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$-characteristic).

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:

  $k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$-characteristic).

- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013b]:

  $k_1 \& k_2$: characteristic $\Leftrightarrow k_1 \otimes k_2$: $\mathcal{I}$-characteristic.

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:
  $$k_1 \& k_2\text{: universal} \Rightarrow k_1 \otimes k_2\text{: universal } (\Rightarrow \mathcal{I}\text{-characteristic}).$$
- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013b]:
  $$k_1 \& k_2\text{: characteristic} \Leftrightarrow k_1 \otimes k_2\text{: } \mathcal{I}\text{-characteristic}.$$

Goal

Extension to $M \geqslant 2$.

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:

$$k_1 \& k_2\text{: universal} \Rightarrow k_1 \otimes k_2\text{: universal } (\Rightarrow \mathcal{I}\text{-characteristic}).$$

- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013b]:

$$k_1 \& k_2\text{: characteristic} \Leftrightarrow k_1 \otimes k_2\text{: } \mathcal{I}\text{-characteristic}.$$

Goal

Extension to $M \geqslant 2$.

Main Challenge

'$\otimes k_m$: $\mathcal{I}$-characteristic $\Leftrightarrow$ k_m: characteristic $(\forall m)$' does NOT hold.

Results [Szabó and Sriperumbudur, 2017]

Proposition (characteristic property)

- $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*
- $\not\Leftarrow$ $[|\mathcal{X}_m| = 2,\ k_m(x,x') = 2\delta_{x,x'} - 1]$

Results [Szabó and Sriperumbudur, 2017]

Proposition (characteristic property)

- $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*
- $\not\Leftarrow$ $[|\mathcal{X}_m| = 2,\ k_m(x,x') = 2\delta_{x,x'} - 1]$

Proposition ($\mathcal{I}$-characteristic property)

- k_1, k_2: *characteristic* $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$*-characteristic.*
- $\Leftarrow$: *for* $\forall M \geqslant 2$.
- k_1, k_2, k_3: *characteristic* $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$*-characteristic [Ex].*

Results [Szabó and Sriperumbudur, 2017]

Proposition (characteristic property)

- $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*
- $\not\Leftarrow$ $[|\mathcal{X}_m| = 2,\ k_m(x, x') = 2\delta_{x,x'} - 1]$

Proposition ($\mathcal{I}$-characteristic property)

- k_1, k_2: *characteristic* $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-*characteristic.*
- $\Leftarrow$: *for* $\forall M \geqslant 2$.
- k_1, k_2, k_3: *characteristic* $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-*characteristic [Ex].*

Proposition ($\mathcal{X}_m = \mathbb{R}^{d_m}$, k_m: continuous, shift-invariant, bounded)

$(k_m)_{m=1}^{M}$-*s are characteristic* $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-*characteristic* $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: *characteristic.*

Results [Szabó and Sriperumbudur, 2017]

Proposition (characteristic property)

- $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*
- $\not\Leftarrow$ $[|\mathcal{X}_m| = 2,\ k_m(x,x') = 2\delta_{x,x'} - 1]$

Proposition ($\mathcal{I}$-characteristic property)

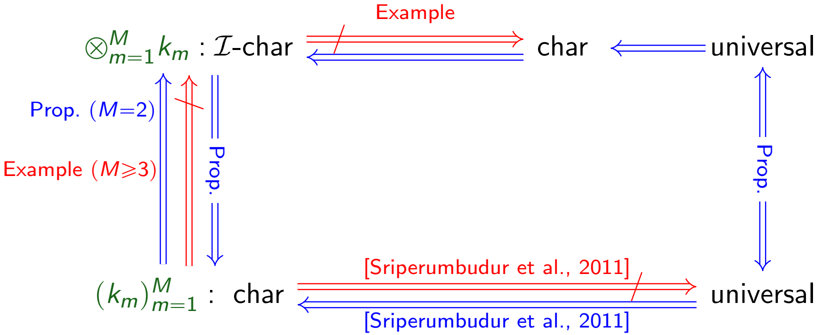
- k_1, k_2: *characteristic* $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-*characteristic.*
- $\Leftarrow$: *for* $\forall M \geqslant 2$.
- k_1, k_2, k_3: *characteristic* $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-*characteristic [Ex].*

Proposition ($\mathcal{X}_m = \mathbb{R}^{d_m}$, k_m: continuous, shift-invariant, bounded)

$(k_m)_{m=1}^{M}$-*s are characteristic* $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-*characteristic* $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: *characteristic.*

Proposition (Universality)

$\otimes_{m=1}^{M} k_m$: *universal* $\Leftrightarrow$ $(k_m)_{m=1}^{M}$ *are universal.*


$\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-char
Example
char
universal
Prop. (M=2)
Example ($M\geqslant3$)
Prop.
Prop.
$(k_m)_{m=1}^{M}$: char
[Sriperumbudur et al., 2011]
universal
[Sriperumbudur et al., 2011]

Covariance operator: finished.

Recall

- KCCA: independence measure,

$$\rho_{\mathrm{KCCA}}(x, y) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)).$$

Recall

- KCCA: independence measure,

$$\rho_{\text{KCCA}}(x, y) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)).$$

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x).$$

Recall

- KCCA: independence measure,

$$\rho_{\text{KCCA}}(x,y) = \sup_{f\in\mathcal{H}_k, g\in\mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)).$$

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\mathrm{d}\mathbb{P}(x).$$

- MMD: (semi)-metric defined by mean embedding,

$$MMD(\mathbb{P},\mathbb{Q}) = \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}.$$

Recall

- KCCA: independence measure,

$$\rho_{\text{KCCA}}(x,y) = \sup_{f\in\mathcal{H}_k, g\in\mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)).$$

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x).$$

- MMD: (semi)-metric defined by mean embedding,

$$MMD(\mathbb{P},\mathbb{Q}) = \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}.$$

- Cross-covariance operator:

$$C_{xy}^c = \mathbb{E}_{xy}\left[\varphi(x)\otimes\psi(y)\right] - \mu_x \otimes \mu_y.$$

Recall

- KCCA: independence measure,

$$\rho_{\text{KCCA}}(x, y) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(x), g(y)).$$

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x).$$

- MMD: (semi)-metric defined by mean embedding,

$$MMD(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

- Cross-covariance operator:

$$C_{xy}^c = \mathbb{E}_{xy}[\varphi(x) \otimes \psi(y)] - \mu_x \otimes \mu_y.$$

- HSIC: independence measure,

$$HSIC(x, y) = \|C_{xy}^c\|_{HS}.$$

No density estimation

Thus,

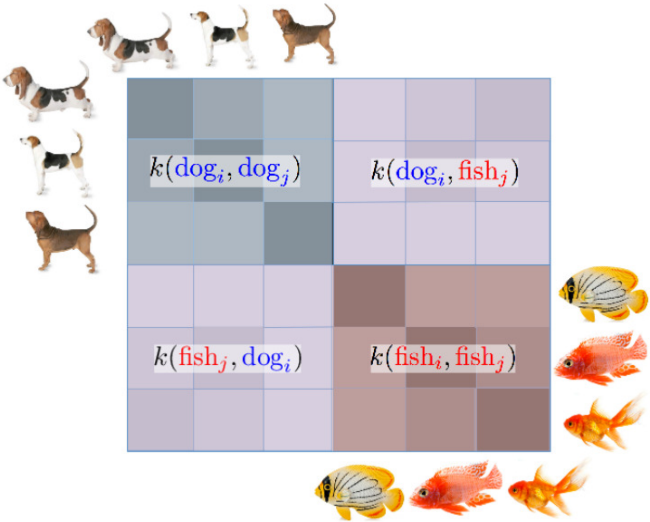
- independence measure,
- distance,
- inner product

measures/estimates on probability distributions

without density estimation!

HSIC estimators

Recall: MMD estimator



$$\widehat{MMD}_u^2(\mathbb{P}, \mathbb{Q}) = \overline{G_{\mathbb{P},\mathbb{P}}} + \overline{G_{\mathbb{Q},\mathbb{Q}}} - 2\overline{G_{\mathbb{P},\mathbb{Q}}}$$ (without diagonals in $\overline{G_{\mathbb{P},\mathbb{P}}}$, $\overline{G_{\mathbb{Q},\mathbb{Q}}}$)

HSIC: intuition. $\mathcal{X}$: images, $\mathcal{Y}$: descriptions.

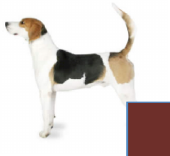
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

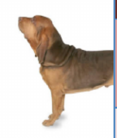
A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose. They need a significant amount of exercise and mental stimulation.

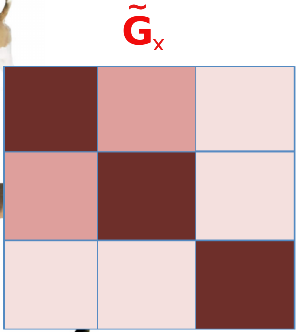
Known for their curiosity, intelligence, and excellent communication skills, the Javanese breed is perfect if you want a responsive, interactive pet, one that will blow in your ear and follow you everywhere.

Text from dogtime.com and petfinder.com

HSIC intuition: Gram matrices



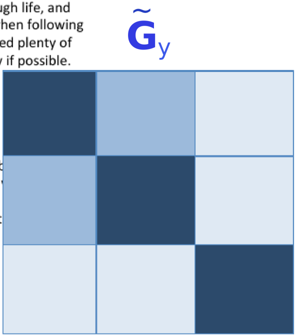




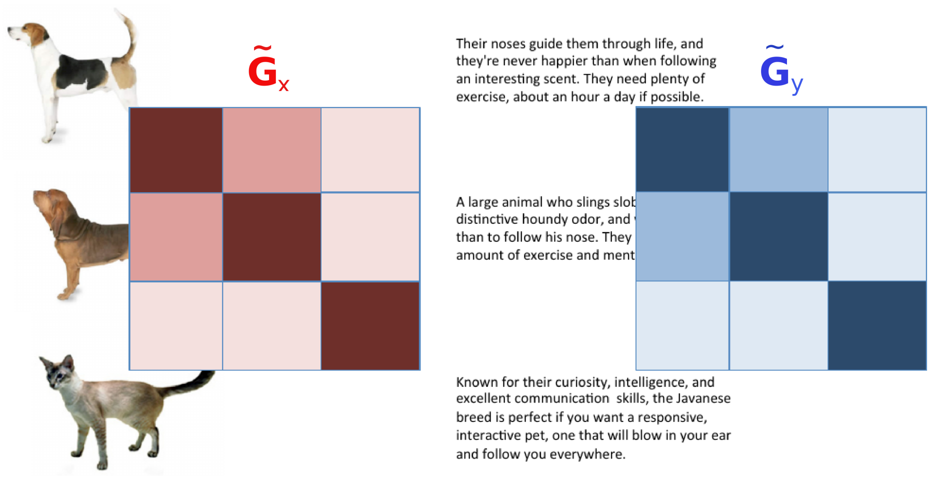
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

A large animal who slings slob
distinctive houndy odor, and
than to follow his nose. They
amount of exercise and ment

Known for their curiosity, intelligence, and excellent communication skills, the Javanese breed is perfect if you want a responsive, interactive pet, one that will blow in your ear and follow you everywhere.



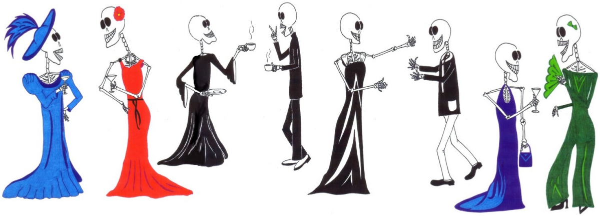
HSIC intuition: Gram matrices



Empirical estimate:

$$\widehat{HSIC^2} = \frac{1}{n^2} \left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F.$$

Cocktail party: HSIC demo



ISA reminder

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right],$$

where $\mathbf{s}^m$-s are non-Gaussian & independent.

- Goal: $\{\mathbf{x}_t\}_{t=1}^T \to \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,

ISA reminder

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right],$$

where $\mathbf{s}^m$-s are non-Gaussian & independent.

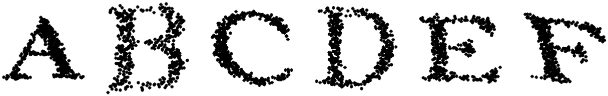
- Goal: $\{\mathbf{x}_t\}_{t=1}^T \rightarrow \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,
- Objective function:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x},$$

$$J(\mathbf{W}) = I\left(\hat{\mathbf{s}}^1, \ldots, \hat{\mathbf{s}}^M\right) \rightarrow \min_{\mathbf{W}}.$$

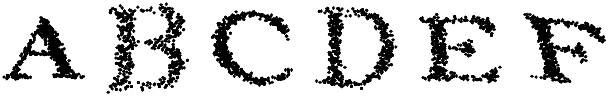
ISA: source, observation

- Hidden sources ($\mathbf{s}$):

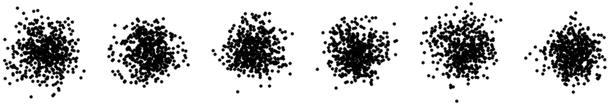


ISA: source, observation

- Hidden sources (**s**):



- Observation (**x**):



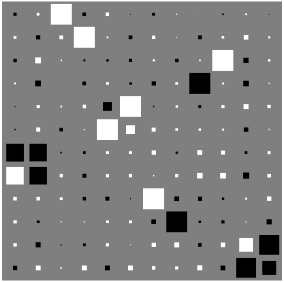
ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

- Performance ($\hat{\mathbf{W}}\mathbf{A}$), ambiguity:

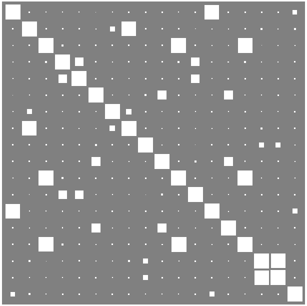


Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation.

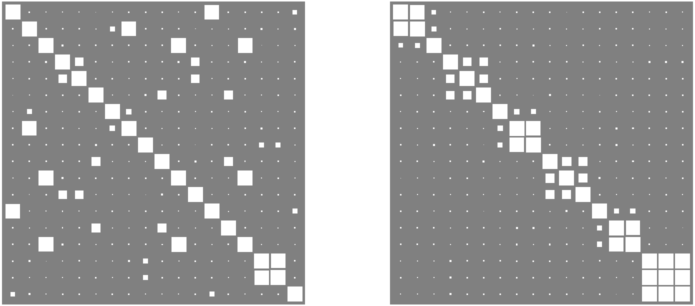
Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$,



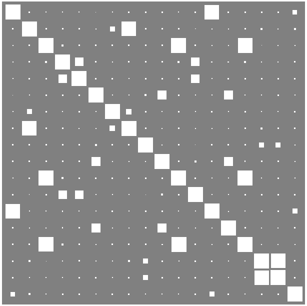
Conjecture: ISA separation theorem [Cardoso, 1998]

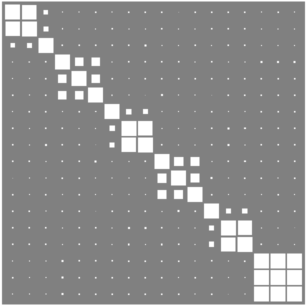
- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$,



Conjecture: ISA separation theorem [Cardoso, 1998]

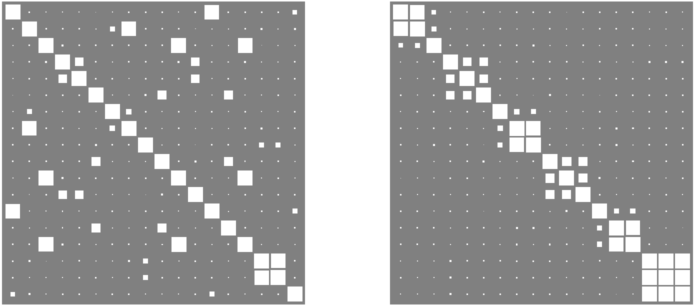
- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$,





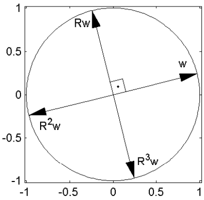
- Basis of the state-of-the-art ISA solvers.

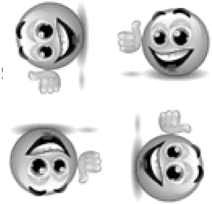
- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$,



- Basis of the state-of-the-art ISA solvers.
- Sufficient conditions [Szabó et al., 2012]:
  - $\mathbf{s}^m$: spherical [Fang et al., 1990].

ISA separation theorem

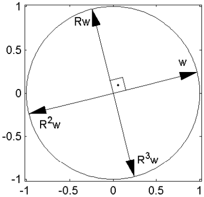


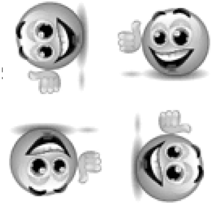


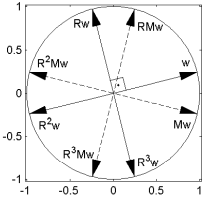
Invariance to

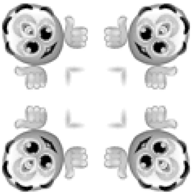
- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.

ISA separation theorem





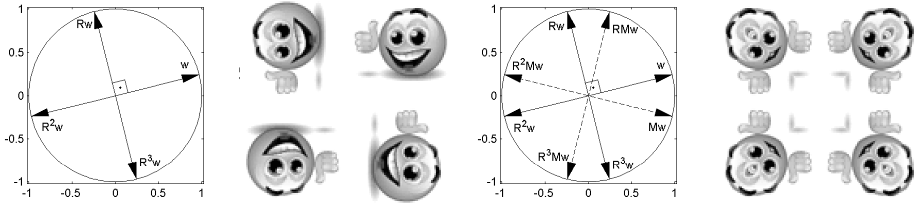




Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.
- permutation and sign: $f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1)$.

ISA separation theorem

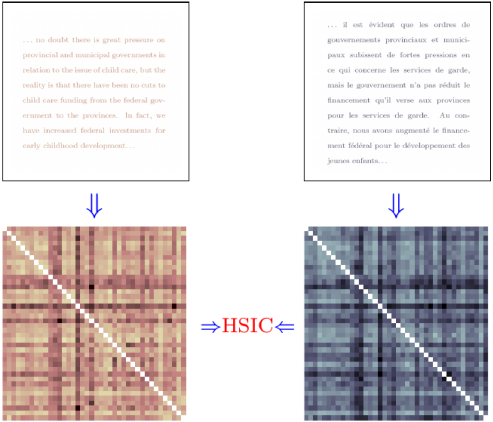


Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.
- permutation and sign: $f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1)$.
- L^p-spherical: $f(u_1, u_2) = h\left(\sum_i |u_i|^p\right) \quad (p > 0)$.

Another HSIC demo: translation

- 5-line extracts.
- representation, kernel: bag-of-words, r-spectrum ($r = 5$).
- sample size: $n = 10$. repetitions: 300.

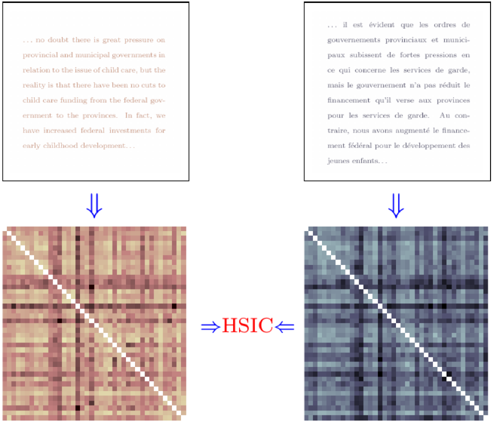


Another HSIC demo: translation

- 5-line extracts.
- representation, kernel: bag-of-words, r-spectrum ($r = 5$).
- sample size: $n = 10$. repetitions: 300.

Results:

- r-spectrum: average Type-II error $= 0$ ($\alpha = 0.05$),
- bag-of-words: 0.18.



Recall: MMD in terms of kernel evaluations

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k} =$$
$$= \mathbb{E}_{x\sim\mathbb{P}, x'\sim\mathbb{P}} k(x, x') + \mathbb{E}_{y\sim\mathbb{Q}, y'\sim\mathbb{Q}} k(y, y')$$
$$- 2\mathbb{E}_{x\sim\mathbb{P}, y\sim\mathbb{Q}} k(x, y).$$

Recall: MMD in terms of kernel evaluations

$$
\begin{aligned}
MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k} &= \\
&= \mathbb{E}_{x\sim\mathbb{P}, x'\sim\mathbb{P}} k(x, x') + \mathbb{E}_{y\sim\mathbb{Q}, y'\sim\mathbb{Q}} k(y, y') \\
&\quad - 2\mathbb{E}_{x\sim\mathbb{P}, y\sim\mathbb{Q}} k(x, y).
\end{aligned}
$$

Question

Can we rewrite HSIC in terms of expected kernel values?

HSIC in terms of kernel evaluations [Gretton et al., 2005a]

$$HSIC^2(x,y) = \left\| C^c_{xy} \right\|^2_{HS} = \left\| C^u_{xy} - \mu_x \otimes \mu_y \right\|^2_{HS}$$

HSIC in terms of kernel evaluations [Gretton et al., 2005a]

$$
\begin{aligned}
HSIC^2(x,y) &= \left\| C^c_{xy} \right\|^2_{HS} = \left\| C^u_{xy} - \mu_x \otimes \mu_y \right\|^2_{HS} \\
&= \left\| C^u_{xy} \right\|^2_{HS} + \left\| \mu_x \otimes \mu_y \right\|^2_{HS} - 2 \left\langle C^u_{xy}, \mu_x \otimes \mu_y \right\rangle_{HS}.
\end{aligned}
$$

HSIC in terms of kernel evaluations [Gretton et al., 2005a]

$$HSIC^2(x,y) = \left\| C_{xy}^c \right\|_{HS}^2 = \left\| C_{xy}^u - \mu_x \otimes \mu_y \right\|_{HS}^2$$
$$= \left\| C_{xy}^u \right\|_{HS}^2 + \left\| \mu_x \otimes \mu_y \right\|_{HS}^2 - 2 \left\langle C_{xy}^u, \mu_x \otimes \mu_y \right\rangle_{HS}.$$

First term:

$$\left\| C_{xy}^u \right\|_{HS}^2 = \left\langle \mathbb{E}_{xy} \left[\varphi(x) \otimes \psi(y) \right], \mathbb{E}_{x'y'} \left[\varphi(x') \otimes \psi(y') \right] \right\rangle_{HS}$$

HSIC in terms of kernel evaluations [Gretton et al., 2005a]

$$HSIC^2(x,y) = \left\|C_{xy}^c\right\|_{HS}^2 = \left\|C_{xy}^u - \mu_x \otimes \mu_y\right\|_{HS}^2$$
$$= \left\|C_{xy}^u\right\|_{HS}^2 + \left\|\mu_x \otimes \mu_y\right\|_{HS}^2 - 2\left\langle C_{xy}^u, \mu_x \otimes \mu_y\right\rangle_{HS}.$$

First term:

$$\left\|C_{xy}^u\right\|_{HS}^2 = \left\langle \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right], \mathbb{E}_{x'y'}\left[\varphi(x') \otimes \psi(y')\right]\right\rangle_{HS}$$
$$= \mathbb{E}_{xy}\mathbb{E}_{x'y'} \underbrace{\left\langle \varphi(x) \otimes \psi(y), \varphi(x') \otimes \psi(y')\right\rangle_{HS}}_{\langle\varphi(x),\varphi(x')\rangle_{\mathcal{H}_k}\langle\psi(y),\psi(y')\rangle_{\mathcal{H}_\ell}}$$

$$\left\langle e_1 \otimes f_1, e_2 \otimes f_2\right\rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)} = \left\langle e_1, e_2\right\rangle_{\mathcal{H}_1}\left\langle f_1, f_2\right\rangle_{\mathcal{H}_2}.$$

HSIC in terms of kernel evaluations [Gretton et al., 2005a]

$$HSIC^2(x,y) = \left\|C_{xy}^c\right\|_{HS}^2 = \left\|C_{xy}^u - \mu_x \otimes \mu_y\right\|_{HS}^2$$
$$= \left\|C_{xy}^u\right\|_{HS}^2 + \left\|\mu_x \otimes \mu_y\right\|_{HS}^2 - 2\left\langle C_{xy}^u, \mu_x \otimes \mu_y\right\rangle_{HS}.$$

First term:

$$\left\|C_{xy}^u\right\|_{HS}^2 = \left\langle \mathbb{E}_{xy}\left[\varphi(x)\otimes\psi(y)\right], \mathbb{E}_{x'y'}\left[\varphi(x')\otimes\psi(y')\right]\right\rangle_{HS}$$
$$= \mathbb{E}_{xy}\mathbb{E}_{x'y'}\underbrace{\left\langle \varphi(x)\otimes\psi(y), \varphi(x')\otimes\psi(y')\right\rangle_{HS}}_{\langle\varphi(x),\varphi(x')\rangle_{\mathcal{H}_k}\langle\psi(y),\psi(y')\rangle_{\mathcal{H}_\ell}}$$
$$= \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y').$$

$$\langle e_1\otimes f_1, e_2\otimes f_2\rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)} = \langle e_1, e_2\rangle_{\mathcal{H}_1}\langle f_1, f_2\rangle_{\mathcal{H}_2}.$$

HSIC: second term

$$\|\mu_x \otimes \mu_y\|_{HS}^2 = \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{HS}$$

HSIC: second term

$$\begin{aligned}\|\mu_x \otimes \mu_y\|_{HS}^2 &= \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{HS} \\ &= \langle \mu_x, \mu_x \rangle_{\mathcal{H}_k} \langle \mu_y, \mu_y \rangle_{\mathcal{H}_\ell}\end{aligned}$$

HSIC: second term

$$\begin{aligned}\|\mu_x \otimes \mu_y\|_{HS}^2 &= \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{HS}\\ &= \langle \mu_x, \mu_x \rangle_{\mathcal{H}_k} \langle \mu_y, \mu_y \rangle_{\mathcal{H}_\ell}\\ &= \mathbb{E}_{xx'} k(x, x') \mathbb{E}_{yy'} \ell(y, y').\end{aligned}$$

HSIC: third term

$$\left\langle C_{xy}^u, \mu_x \otimes \mu_y \right\rangle_{HS} = \langle \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right], \mu_x \otimes \mu_y \rangle_{HS}$$

HSIC: third term

$$\left\langle C_{xy}^{u}, \mu_x \otimes \mu_y \right\rangle_{HS} = \langle \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right], \mu_x \otimes \mu_y \rangle_{HS}$$

$$= \mathbb{E}_{xy} \underbrace{\langle \varphi(x) \otimes \psi(y), \mu_x \otimes \mu_y \rangle_{HS}}_{\underbrace{\langle \varphi(x), \mu_x \rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x'} k(x,x')} \underbrace{\langle \psi(y), \mu_y \rangle_{\mathcal{H}_\ell}}_{\mathbb{E}_{y'} \ell(y,y')}}$$

HSIC: third term

$$
\begin{aligned}
\left\langle C_{xy}^{u}, \mu_x \otimes \mu_y \right\rangle_{HS} &= \left\langle \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right], \mu_x \otimes \mu_y \right\rangle_{HS} \\
&= \mathbb{E}_{xy} \underbrace{\left\langle \varphi(x) \otimes \psi(y), \mu_x \otimes \mu_y \right\rangle_{HS}}_{\underbrace{\langle \varphi(x), \mu_x \rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x'} k(x,x')} \underbrace{\langle \psi(y), \mu_y \rangle_{\mathcal{H}_\ell}}_{\mathbb{E}_{y'} \ell(y,y')}} \\
&= \mathbb{E}_{xy}\left[\mathbb{E}_{x'} k(x,x') \mathbb{E}_{y'} \ell(y,y')\right].
\end{aligned}
$$

HSIC: after gathering the terms

$$
\begin{aligned}
HSIC^2(x,y) &= \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') \\
&\quad - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right]. \\
&=: a + b - 2c.
\end{aligned}
$$

HSIC: after gathering the terms

$$
\begin{aligned}
HSIC^2(x,y) &= \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') \\
&\quad - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right]. \\
&=: a + b - 2c.
\end{aligned}
$$

Idea: given $\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{xy}$,

- Let us estimate C_{xy}^u, μ_x, μ_y empirically.

HSIC: after gathering the terms

$$HSIC^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') \\ - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right]. \\ =: a + b - 2c.$$

Idea: given $\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{xy}$,

- Let us estimate C_{xy}^u, μ_x, μ_y empirically.

Result

$\widehat{HSIC_b^2}(x,y) = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F$: see the intuition.

First term:

$$a = \left\| C_{xy}^u \right\|_{HS}^2 = \mathbb{E}_{xy}\mathbb{E}_{x'y'} k(x,x')\ell(y,y'),$$

$$\hat{a} = \left\| \widehat{C_{xy}^u} \right\|_{HS}^2 =$$

HSIC estimation: from $\widehat{C^u_{xy}}, \hat{\mu}_x, \hat{\mu}_y$

First term:

$$a = \left\| C^u_{xy} \right\|^2_{HS} = \mathbb{E}_{xy}\mathbb{E}_{x'y'} k(x,x')\ell(y,y'),$$

$$\hat{a} = \left\| \widehat{C^u_{xy}} \right\|^2_{HS} = \left\langle \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i) \otimes \psi(y_i), \frac{1}{n}\sum_{j=1}^{n} \varphi(x_j) \otimes \psi(y_j) \right\rangle_{HS}$$

HSIC estimation: from $\widehat{C^u_{xy}}, \hat{\mu}_x, \hat{\mu}_y$

First term:

$$a = \left\| C^u_{xy} \right\|^2_{HS} = \mathbb{E}_{xy}\mathbb{E}_{x'y'} k(x,x')\ell(y,y'),$$

$$\hat{a} = \left\| \widehat{C^u_{xy}} \right\|^2_{HS} = \left\langle \frac{1}{n}\sum_{i=1}^n \varphi(x_i) \otimes \psi(y_i), \frac{1}{n}\sum_{j=1}^n \varphi(x_j) \otimes \psi(y_j) \right\rangle_{HS}$$

$$= \frac{1}{n^2} \sum_{i,j=1}^n (\mathbf{G}_x)_{ij} (\mathbf{G}_y)_{ij}$$

HSIC estimation: from $\widehat{C_{xy}^u}, \hat{\mu}_x, \hat{\mu}_y$

First term:

$$a = \left\| C_{xy}^u \right\|_{HS}^2 = \mathbb{E}_{xy}\mathbb{E}_{x'y'} k(x,x')\ell(y,y'),$$

$$\hat{a} = \left\| \widehat{C_{xy}^u} \right\|_{HS}^2 = \left\langle \frac{1}{n}\sum_{i=1}^n \varphi(x_i)\otimes\psi(y_i), \frac{1}{n}\sum_{j=1}^n \varphi(x_j)\otimes\psi(y_j) \right\rangle_{HS}$$

$$= \frac{1}{n^2}\sum_{i,j=1}^n (\mathbf{G}_x)_{ij}(\mathbf{G}_y)_{ij} = \frac{1}{n^2}\left\langle \mathbf{G}_x, \mathbf{G}_y \right\rangle_F = \frac{1}{n^2}\operatorname{tr}(\mathbf{G}_x\mathbf{G}_y).$$

HSIC estimation: 2nd term

$$b = \|\mu_x \otimes \mu_y\|_{HS}^2 = \mathbb{E}_{xx'} k(x, x') \mathbb{E}_{yy'} \ell(y, y').$$

$$\hat{b} = \|\hat{\mu}_x \otimes \hat{\mu}_y\|_{HS}^2$$

HSIC estimation: 2nd term

$b = \|\mu_x \otimes \mu_y\|_{HS}^2 = \mathbb{E}_{xx'} k(x, x') \mathbb{E}_{yy'} \ell(y, y')$.

$\hat{b} = \|\hat{\mu}_x \otimes \hat{\mu}_y\|_{HS}^2 = \langle \hat{\mu}_x \otimes \hat{\mu}_y, \hat{\mu}_x \otimes \hat{\mu}_y \rangle_{HS}$

HSIC estimation: 2nd term

$$b = \|\mu_x \otimes \mu_y\|_{HS}^2 = \mathbb{E}_{xx'} k(x, x') \mathbb{E}_{yy'} \ell(y, y').$$

$$\hat{b} = \|\hat{\mu}_x \otimes \hat{\mu}_y\|_{HS}^2 = \langle \hat{\mu}_x \otimes \hat{\mu}_y, \hat{\mu}_x \otimes \hat{\mu}_y \rangle_{HS}$$

$$= \left\langle \left[\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)\right] \otimes \left[\frac{1}{n} \sum_{j=1}^{n} \psi(y_j)\right], \left[\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)\right] \otimes \left[\frac{1}{n} \sum_{j=1}^{n} \psi(y_j)\right] \right\rangle_{HS}$$

HSIC estimation: 2nd term

$$b = \|\mu_x \otimes \mu_y\|_{HS}^2 = \mathbb{E}_{xx'} k(x,x') \mathbb{E}_{yy'} \ell(y,y').$$

$$\hat{b} = \|\hat{\mu}_x \otimes \hat{\mu}_y\|_{HS}^2 = \langle \hat{\mu}_x \otimes \hat{\mu}_y, \hat{\mu}_x \otimes \hat{\mu}_y \rangle_{HS}$$

$$= \left\langle \left[\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\right] \otimes \left[\frac{1}{n}\sum_{j=1}^n \psi(y_j)\right], \left[\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\right] \otimes \left[\frac{1}{n}\sum_{j=1}^n \psi(y_j)\right] \right\rangle_{HS}$$

$$= \left[\frac{1}{n^2}\sum_{i,j=1}^n k(x_i,x_j)\right] \left[\frac{1}{n^2}\sum_{i,j=1}^n \ell(x_i,x_j)\right]$$

HSIC estimation: 2nd term

$$b = \|\mu_x \otimes \mu_y\|_{HS}^2 = \mathbb{E}_{xx'} k(x,x') \mathbb{E}_{yy'} \ell(y,y').$$

$$\hat{b} = \|\hat{\mu}_x \otimes \hat{\mu}_y\|_{HS}^2 = \langle \hat{\mu}_x \otimes \hat{\mu}_y, \hat{\mu}_x \otimes \hat{\mu}_y \rangle_{HS}$$

$$= \left\langle \left[\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\right] \otimes \left[\frac{1}{n}\sum_{j=1}^n \psi(y_j)\right], \left[\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\right] \otimes \left[\frac{1}{n}\sum_{j=1}^n \psi(y_j)\right] \right\rangle_{HS}$$

$$= \left[\frac{1}{n^2}\sum_{i,j=1}^n k(x_i,x_j)\right]\left[\frac{1}{n^2}\sum_{i,j=1}^n \ell(x_i,x_j)\right] = \frac{1}{n^4}\left(\mathbf{1}^T \mathbf{G}_x \mathbf{1}\right)\left(\mathbf{1}^T \mathbf{G}_y \mathbf{1}\right).$$

HSIC estimation: 3rd term (without '−2')

$$c = \left\langle C_{xy}^u, \mu_x \otimes \mu_y \right\rangle_{HS},$$

$$\hat{c} = \left\langle \widehat{C_{xy}^u}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{HS}$$

HSIC estimation: 3rd term (without '−2')

$$c = \left\langle C^u_{xy}, \mu_x \otimes \mu_y \right\rangle_{HS},$$

$$\hat{c} = \left\langle \widehat{C^u_{xy}}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{HS}$$

$$= \left\langle \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^{n} \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^{n} \psi(y_b)\right] \right\rangle_{HS}$$

HSIC estimation: 3rd term (without '−2')

$$c = \left\langle C^u_{xy}, \mu_x \otimes \mu_y \right\rangle_{HS},$$

$$\hat{c} = \left\langle \widehat{C^u_{xy}}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{HS}$$

$$= \left\langle \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^{n} \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^{n} \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left\langle \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^{n} \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^{n} \psi(y_b)\right] \right\rangle_{HS}$$

HSIC estimation: 3rd term (without '−2')

$$c = \left\langle C_{xy}^u, \mu_x \otimes \mu_y \right\rangle_{HS},$$

$$\hat{c} = \left\langle \widehat{C_{xy}^u}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{HS}$$

$$= \left\langle \frac{1}{n}\sum_{i=1}^n \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^n \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^n \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^n \left\langle \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^n \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^n \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^n \underbrace{\left\langle \varphi(x_i), \frac{1}{n}\sum_{a=1}^n \varphi(x_a) \right\rangle_{\mathcal{H}_k}}_{\frac{1}{n}\sum_{a=1}^n k(x_i,x_a)} \underbrace{\left\langle \psi(y_i), \frac{1}{n}\sum_{b=1}^n \psi(y_b) \right\rangle_{\mathcal{H}_\ell}}_{\frac{1}{n}\sum_{b=1}^n \ell(y_i,y_b)}$$

HSIC estimation: 3rd term (without '−2')

$$c = \left\langle C^u_{xy}, \mu_x \otimes \mu_y \right\rangle_{HS},$$

$$\hat{c} = \left\langle \widehat{C^u_{xy}}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{HS}$$

$$= \left\langle \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^{n} \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^{n} \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left\langle \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^{n} \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^{n} \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left\langle \varphi(x_i), \frac{1}{n}\sum_{a=1}^{n} \varphi(x_a) \right\rangle_{\mathcal{H}_k}}_{\frac{1}{n}\sum_{a=1}^{n} k(x_i,x_a)} \underbrace{\left\langle \psi(y_i), \frac{1}{n}\sum_{b=1}^{n} \psi(y_b) \right\rangle_{\mathcal{H}_\ell}}_{\frac{1}{n}\sum_{b=1}^{n} \ell(y_i,y_b)}$$

$$= \frac{1}{n^3} \sum_{a,b=1}^{n} \underbrace{\left[\sum_{i=1}^{n} k(x_i,x_a)\ell(y_i,y_b)\right]}_{(\mathbf{G}_x\mathbf{G}_y)_{a,b}}$$

HSIC estimation: 3rd term (without '−2')

$$c = \left\langle C^u_{xy}, \mu_x \otimes \mu_y \right\rangle_{HS},$$

$$\hat{c} = \left\langle \widehat{C^u_{xy}}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{HS}$$

$$= \left\langle \frac{1}{n}\sum_{i=1}^n \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^n \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^n \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^n \left\langle \varphi(x_i) \otimes \psi(y_i), \left[\frac{1}{n}\sum_{a=1}^n \varphi(x_a)\right] \otimes \left[\frac{1}{n}\sum_{b=1}^n \psi(y_b)\right] \right\rangle_{HS}$$

$$= \frac{1}{n}\sum_{i=1}^n \underbrace{\left\langle \varphi(x_i), \frac{1}{n}\sum_{a=1}^n \varphi(x_a) \right\rangle_{\mathcal{H}_k}}_{\frac{1}{n}\sum_{a=1}^n k(x_i,x_a)} \underbrace{\left\langle \psi(y_i), \frac{1}{n}\sum_{b=1}^n \psi(y_b) \right\rangle_{\mathcal{H}_\ell}}_{\frac{1}{n}\sum_{b=1}^n \ell(y_i,y_b)}$$

$$= \frac{1}{n^3}\sum_{a,b=1}^n \underbrace{\left[\sum_{i=1}^n k(x_i,x_a)\ell(y_i,y_b)\right]}_{(\mathbf{G}_x\mathbf{G}_y)_{a,b}} = \frac{1}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1}.$$

HSIC estimation: putting together

$$\widehat{HSIC_b^2}(x, y) =: \hat{a} + \hat{b} - 2\hat{c}$$

HSIC estimation: putting together

$$\widehat{HSIC_b^2}(x,y) =: \hat{a} + \hat{b} - 2\hat{c}$$
$$= \frac{1}{n^2}\operatorname{tr}(\mathbf{G}_x\mathbf{G}_y) + \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{G}_x\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{G}_y\mathbf{1}\right) - \frac{2}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1}$$

HSIC estimation: putting together

$$\widehat{HSIC}_b^2(x,y) =: \hat{a} + \hat{b} - 2\hat{c}$$
$$= \frac{1}{n^2}\operatorname{tr}(\mathbf{G}_x\mathbf{G}_y) + \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{G}_x\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{G}_y\mathbf{1}\right) - \frac{2}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1}$$
$$= \frac{1}{n^2}\operatorname{tr}\Bigg(\underbrace{\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{1}\mathbf{1}^T\mathbf{G}_y}_{\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_x\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_y}\Bigg)$$

HSIC estimation: putting together

$$\widehat{HSIC_b^2}(x,y) =: \hat{a} + \hat{b} - 2\hat{c}$$
$$= \frac{1}{n^2}\mathrm{tr}(\mathbf{G}_x\mathbf{G}_y) + \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{G}_x\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{G}_y\mathbf{1}\right) - \frac{2}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1}$$
$$= \frac{1}{n^2}\mathrm{tr}\Bigg(\underbrace{\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{1}\mathbf{1}^T\mathbf{G}_y}_{\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_x\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_y}\Bigg)$$
$$= \frac{1}{n^2}\mathrm{tr}\left(\mathbf{H}\mathbf{G}_x\mathbf{H}\mathbf{G}_y\right)$$

HSIC estimation: putting together

$$\widehat{HSIC_b^2}(x,y) =: \hat{a} + \hat{b} - 2\hat{c}$$

$$= \frac{1}{n^2}\,\mathrm{tr}(\mathbf{G}_x\mathbf{G}_y) + \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{G}_x\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{G}_y\mathbf{1}\right) - \frac{2}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1}$$

$$= \frac{1}{n^2}\,\mathrm{tr}\left(\underbrace{\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{1}\mathbf{1}^T\mathbf{G}_y}_{\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_x\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_y}\right)$$

$$= \frac{1}{n^2}\,\mathrm{tr}\left(\mathbf{H}\mathbf{G}_x\mathbf{H}\mathbf{G}_y\right) = \frac{1}{n^2}\,\mathrm{tr}\left(\underbrace{\mathbf{H}\mathbf{G}_x\mathbf{H}}_{\tilde{\mathbf{G}}_x}\underbrace{\mathbf{H}\mathbf{G}_y\mathbf{H}}_{\tilde{\mathbf{G}}_y}\right)$$

HSIC estimation: putting together

$$
\begin{aligned}
\widehat{HSIC_b^2}(x,y) &=: \hat{a} + \hat{b} - 2\hat{c} \\
&= \frac{1}{n^2}\operatorname{tr}(\mathbf{G}_x\mathbf{G}_y) + \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{G}_x\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{G}_y\mathbf{1}\right) - \frac{2}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1} \\
&= \frac{1}{n^2}\operatorname{tr}\Bigg(\underbrace{\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{1}\mathbf{1}^T\mathbf{G}_y}_{\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_x\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_y}\Bigg) \\
&= \frac{1}{n^2}\operatorname{tr}\left(\mathbf{H}\mathbf{G}_x\mathbf{H}\mathbf{G}_y\right) = \frac{1}{n^2}\operatorname{tr}\Bigg(\underbrace{\mathbf{H}\mathbf{G}_x\mathbf{H}}_{\tilde{\mathbf{G}}_x}\underbrace{\mathbf{H}\mathbf{G}_y\mathbf{H}}_{\tilde{\mathbf{G}}_y}\Bigg) = \frac{1}{n^2}\left\langle\tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y\right\rangle_F.
\end{aligned}
$$

HSIC estimation: putting together

$$
\begin{aligned}
\widehat{HSIC_b^2}(x,y) &=: \hat{a} + \hat{b} - 2\hat{c} \\
&= \frac{1}{n^2}\operatorname{tr}(\mathbf{G}_x\mathbf{G}_y) + \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{G}_x\mathbf{1}\right)\left(\mathbf{1}^T\mathbf{G}_y\mathbf{1}\right) - \frac{2}{n^3}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y\mathbf{1} \\
&= \frac{1}{n^2}\operatorname{tr}\Bigg(\underbrace{\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{G}_y + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{G}_x\mathbf{1}\mathbf{1}^T\mathbf{G}_y}_{\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_x\left(\mathbf{I}_n - \frac{\mathbf{E}_n}{n}\right)\mathbf{G}_y}\Bigg) \\
&= \frac{1}{n^2}\operatorname{tr}\left(\mathbf{H}\mathbf{G}_x\mathbf{H}\mathbf{G}_y\right) = \frac{1}{n^2}\operatorname{tr}\Bigg(\underbrace{\mathbf{H}\mathbf{G}_x\mathbf{H}}_{\tilde{\mathbf{G}}_x}\underbrace{\mathbf{H}\mathbf{G}_y\mathbf{H}}_{\tilde{\mathbf{G}}_y}\Bigg) = \frac{1}{n^2}\left\langle\tilde{\mathbf{G}}_x,\tilde{\mathbf{G}}_y\right\rangle_F.
\end{aligned}
$$

Bias: $\mathcal{O}\left(\frac{1}{n}\right)$.

Reminder: MMD^2, $\widehat{MMD}_b^2$, $\widehat{MMD}_u^2$

$$MMD^2(\mathbb{P},\mathbb{Q}) := \mathbb{E}_{xx'}k(x,x') + \mathbb{E}_{yy'}k(y,y') - 2\mathbb{E}_{xy}k(x,y),$$

$$\widehat{MMD}_b^2(\mathbb{P},\mathbb{Q}) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i,x_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j),$$

$$\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j).$$

$\widehat{HSIC_b^2}$ until now

$$\begin{aligned} HSIC^2(x,y) &= \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') \\ &\quad - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right], \\ \widehat{HSIC_b^2}(x,y) &= \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j)\ell(y_i,y_j) + \ldots. \end{aligned}$$

$\widehat{HSIC_b^2}$ until now

$$HSIC^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right],$$

$$\widehat{HSIC_b^2}(x,y) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j)\ell(y_i,y_j) + \ldots.$$

- x, x' should be independent, but
- with plug-in: $i = j$, it introduces bias.

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij}\ell_{ij},$$

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij}\ell_{ij}, \qquad \hat{a}_u = \frac{1}{n(n-1)} \sum_{i \neq j} k_{ij}\ell_{ij}$$

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij}\ell_{ij}, \qquad \hat{a}_u = \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j} k_{ij}\ell_{ij}}_{\frac{1}{(n)_2} \sum_{(i,j)\in I_2^n} k_{ij}\ell_{ij}},$$

$I_p^n = \{(i_1, \ldots, i_p) : i_j \in \{1, \ldots, n\}$ *without* replacement$\}$, $(n)_p = \left|I_p^n\right|$.

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij}\ell_{ij}, \qquad \hat{a}_u = \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j} k_{ij}\ell_{ij}}_{\frac{1}{(n)_2} \sum_{(i,j) \in I_2^n} k_{ij}\ell_{ij}},$$

$$\hat{c}_b = \frac{1}{n^3} \sum_{i,q,r=1}^{n} k_{iq}\ell_{ir},$$

$I_p^n = \{(i_1, \ldots, i_p) : i_j \in \{1, \ldots, n\}$ *without* replacement$\}$, $(n)_p = \left|I_p^n\right|$.

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}\ell_{ij}, \qquad \hat{a}_u = \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j} k_{ij}\ell_{ij}}_{\frac{1}{(n)_2}\sum_{(i,j)\in I_2^n} k_{ij}\ell_{ij}},$$

$$\hat{c}_b = \frac{1}{n^3}\sum_{i,q,r=1}^{n} k_{iq}\ell_{ir}, \qquad \hat{c}_u = \frac{1}{(n)_3}\sum_{(i,q,r)\in I_3^n} k_{iq}\ell_{ir},$$

$I_p^n = \{(i_1, \ldots, i_p) : i_j \in \{1, \ldots, n\}$ *without* replacement$\}$, $(n)_p = \left|I_p^n\right|$.

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij}\ell_{ij}, \qquad \hat{a}_u = \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j} k_{ij}\ell_{ij}}_{\frac{1}{(n)_2} \sum_{(i,j) \in I_2^n} k_{ij}\ell_{ij}},$$

$$\hat{c}_b = \frac{1}{n^3} \sum_{i,q,r=1}^{n} k_{iq}\ell_{ir}, \qquad \hat{c}_u = \frac{1}{(n)_3} \sum_{(i,q,r) \in I_3^n} k_{iq}\ell_{ir},$$

$$\hat{b}_b = \frac{1}{n^4} \sum_{i,j,q,r=1}^{n} k_{ij}\ell_{qr},$$

$I_p^n = \{(i_1, \ldots, i_p) : i_j \in \{1, \ldots, n\}$ *without* replacement$\}$, $(n)_p = \left|I_p^n\right|$.

HSIC: unbiased estimator

Idea: get rid of the $i = j$-type terms. Let $k_{ij} := k(x_i, x_j)$, $\ell_{ij} := \ell(y_i, y_j)$.

$$\hat{a}_b = \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij}\ell_{ij}, \qquad \hat{a}_u = \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j} k_{ij}\ell_{ij}}_{\frac{1}{(n)_2} \sum_{(i,j) \in I_2^n} k_{ij}\ell_{ij}},$$

$$\hat{c}_b = \frac{1}{n^3} \sum_{i,q,r=1}^{n} k_{iq}\ell_{ir}, \qquad \hat{c}_u = \frac{1}{(n)_3} \sum_{(i,q,r) \in I_3^n} k_{iq}\ell_{ir},$$

$$\hat{b}_b = \frac{1}{n^4} \sum_{i,j,q,r=1}^{n} k_{ij}\ell_{qr}, \qquad \hat{b}_u = \frac{1}{(n)_4} \sum_{(i,j,q,r) \in I_4^n} k_{ij}\ell_{qr}.$$

$I_p^n = \{(i_1, \ldots, i_p) : i_j \in \{1, \ldots, n\}$ *without* replacement$\}$, $(n)_p = \left|I_p^n\right|$.

HSIC: resulting unbiased estimator

After some linear algebra [Gretton et al., 2005a], $(M)_{++} := \sum_{i,j} M_{ij}$,

$$\widehat{HSIC_b^2}(x,y) = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F,$$

$$\widehat{HSIC_u^2}(x,y) = \frac{1}{n(n-3)}\left[\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F - \frac{2}{n-2}(\tilde{\mathbf{G}}_x\tilde{\mathbf{G}}_y)_{++} + \frac{1}{(n-1)(n-2)}(\tilde{\mathbf{G}}_x)_{++}(\tilde{\mathbf{G}}_y)_{++}\right].$$

Estimation in practice: few ITE examples

KCCA estimation: Matlab

Goal: estimate KCCA,

```
>ds = [2;3;4]; Y = rand(sum(ds),5000);
>mult = 1
>co = IKCCA_initialization(mult);
>KCCA = IKCCA_estimation(Y,ds,co);
```

KCCA estimation: Matlab

Goal: estimate KCCA,

```
>ds = [2;3;4]; Y = rand(sum(ds),5000);
>mult = 1
>co = IKCCA_initialization(mult);
>KCCA = IKCCA_estimation(Y,ds,co);
```

Alternative initialization:

```
>co = IKCCA_initialization(mult,{'kappa',0.01,'eta',0.001});
```

where κ: regularization constant, η: low-rank approximation.

KCCA & HSIC estimation: Matlab

Goal: estimate KCCA,

```
>ds = [2;3;4]; Y = rand(sum(ds),5000);
>mult = 1
>co = IKCCA_initialization(mult);
>KCCA = IKCCA_estimation(Y,ds,co);
```

Alternative initialization:

```
>co = IKCCA_initialization(mult,{'kappa',0.01,'eta',0.001});
```

where κ: regularization constant, η: low-rank approximation.

Note: HSIC similarly.

MMD estimation: Matlab

Using for example U-statistic:

```
>X1 = randn(3,2000); X2 = randn(3,3000);
>mult = 1;
>co = DMMD_Ustat_initialization(mult);
>MMD = DMMD_Ustat_estimation(X1,X2,co);
```

MMD estimation: Matlab

Using for example U-statistic:

```
>X1 = randn(3,2000); X2 = randn(3,3000);
>mult = 1;
>co = DMMD_Ustat_initialization(mult);
>MMD = DMMD_Ustat_estimation(X1,X2,co);
```

With low-rank approximation, and setting some parameters:

```
co2 = DMMD_Ustat_iChol_initialization(mult)
co3 = DMMD_Ustat_iChol_initialization(mult,{'sigma',0.2,
'eta',0.01})
```

HSIC estimation: Python

Import ITE (1x), generate observations:

```
>>> import ite
>>> from numpy.random import randn
>>> from numpy import array
>>> ds = array([2, 3, 4])
>>> t = 1000
>>> y = randn(t, sum(ds))
```

HSIC estimation: Python

Import ITE (1x), generate observations:

```
>>> import ite
>>> from numpy.random import randn
>>> from numpy import array
>>> ds = array([2, 3, 4])
>>> t = 1000
>>> y = randn(t, sum(ds))
```

Estimate HSIC:

```
>>> co = ite.cost.BIHSIC_IChol()
>>> hsic = co.estimation(y, ds)
```

Alternative initialization-1:

```
>>> co2 = ite.cost.BIHSIC_IChol(eta=1e-3)
>>> hsic2 = co2.estimation(y, ds)
```

HSIC estimation: Python

Alternative initialization-1:

```
>>> co2 = ite.cost.BIHSIC_IChol(eta=1e-3)
>>> hsic2 = co2.estimation(y, ds)
```

Alternative-2:

```
>>> from ite.cost.x_kernel import Kernel
>>> k = Kernel({'name': 'RBF','sigma': 1})
>>> co3 = ite.cost.BIHSIC_IChol(kernel=k, eta=1e-3)
>>> hsic3 = co3.estimation(y, ds)
```

HSIC & KCCA estimation: Python

Alternative initialization-1:

```
>>> co2 = ite.cost.BIHSIC_IChol(eta=1e-3)
>>> hsic2 = co2.estimation(y, ds)
```

Alternative-2:

```
>>> from ite.cost.x_kernel import Kernel
>>> k = Kernel({'name': 'RBF','sigma': 1})
>>> co3 = ite.cost.BIHSIC_IChol(kernel=k, eta=1e-3)
>>> hsic3 = co3.estimation(y, ds)
```

Note: KCCA similarly.

MMD estimation: Python

Import ITE, generate observations:

```
>>> import ite
>>> from numpy.random import randn
>>> dim = 3
>>> t1, t2 = 2000, 3000
>>> y1 = randn(t1, dim)
>>> y2 = randn(t2, dim)
```

MMD estimation: Python

Import ITE, generate observations:

```
>>> import ite
>>> from numpy.random import randn
>>> dim = 3
>>> t1, t2 = 2000, 3000
>>> y1 = randn(t1, dim)
>>> y2 = randn(t2, dim)
```

Estimate MMD:

```
>>> co = ite.cost.BDMMD_UStat_IChol()
>>> mmd = co.estimation(y1, y2)
```

MMD estimation: Python

Alternative initialization-1:

```
>>> co2 = ite.cost.BDMMD_UStat_IChol(eta=1e-2)
>>> mmd2 = co2.estimation(y1, y2)
```

MMD estimation: Python

Alternative initialization-1:

```
>>> co2 = ite.cost.BDMMD_UStat_IChol(eta=1e-2)
>>> mmd2 = co2.estimation(y1, y2)
```

Alternative-2:

```
>>> k = Kernel('name': 'RBF','sigma': 1)
>>> co3 = ite.cost.BDMMD_UStat_IChol(kernel=k,eta=1e-2)
>>> mmd3 = co3.estimation(y1, y2)
```

Towards unbiased estimators

- MMD, HSIC: $\mathbb{E}_{x,x'}k(x,x')$-type quantities.

Towards unbiased estimators

- MMD, HSIC: $\mathbb{E}_{x,x'} k(x, x')$-type quantities.
- x,x': independence.

Towards unbiased estimators

- MMD, HSIC: $\mathbb{E}_{x,x'} k(x, x')$-type quantities.
- x,x': independence.
- Plugin methods: $i = j$, biased.

Towards unbiased estimators

- MMD, HSIC: $\mathbb{E}_{x,x'} k(x, x')$-type quantities.
- x,x': independence.
- Plugin methods: $i = j$, biased.
- If we restrict to $i \neq j$, we got unbiased estimators.

Towards unbiased estimators

- MMD, HSIC: $\mathbb{E}_{x,x'}k(x,x')$-type quantities.
- x,x': independence.
- Plugin methods: $i = j$, biased.
- If we restrict to $i \neq j$, we got unbiased estimators.

Question

What is happening here? Concentration of the estimators?
$\rightarrow$ hypothesis testing: our statistics := these estimators

Unbiased estimators for $\mathbb{E}_{x,x'}k(x,x')$-type quantities – extensions of average

Task

- Goal: estimate

$$\theta(\mathbb{P}) := \mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m).$$

Task

- Goal: estimate

$$\theta(\mathbb{P}) := \mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m).$$

- Given: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathbb{P}$, $n \geqslant m$.

Task

- Goal: estimate

$$\theta(\mathbb{P}) := \mathbb{E}_{\mathbb{P}} h\left(X_1, \ldots, X_m\right).$$

- Given: $x_1, \ldots, x_n \stackrel{i.i.d.}{\sim} \mathbb{P}$, $n \geqslant m$.
- Assume (w.l.o.g.): h is symmetric,

$$h(x_1, \ldots, x_m) = h\left(x_{\pi(1)}, \ldots, x_{\pi(m)}\right) \ \forall \pi \text{ permutation.}$$

Example: $k(x, x') = k(x', x)$.

Task

- Goal: estimate

$$\theta(\mathbb{P}) := \mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m).$$

- Given: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathbb{P}$, $n \geqslant m$.
- Assume (w.l.o.g.): h is symmetric,

$$h(x_1, \ldots, x_m) = h\left(x_{\pi(1)}, \ldots, x_{\pi(m)}\right) \ \forall \pi \text{ permutation.}$$

  Example: $k(x, x') = k(x', x)$.
- Otherwise: $h \leftarrow \frac{1}{m!} \sum_{\pi} h\left(x_{\pi(1)}, \ldots, x_{\pi(m)}\right)$.

U-statistic

- Estimator for $\mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m)$:

$$U_n = U(x_1, \ldots, x_n) = \frac{1}{\binom{n}{m}} \sum_c h\left(x_{i_1}, \ldots, x_{i_m}\right),$$

$\sum_c$: m-tuples without replacement.

U-statistic

- Estimator for $\mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m)$:

$$U_n = U(x_1, \ldots, x_n) = \frac{1}{\binom{n}{m}} \sum_c h\left(x_{i_1}, \ldots, x_{i_m}\right),$$

  $\sum_c$: m-tuples without replacement.
- U_n: unbiased, i.e. $\mathbb{E}_{\mathbb{P}}(U_n) = \theta$.

V-statistic

- Estimator for $\mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m)$:

$$V_n = V(x_1, \ldots, x_n) = \frac{1}{n^m} \sum_{i_1=1}^{m} \ldots \sum_{i_m=1}^{n} h\left(x_{i_1}, \ldots, x_{i_m}\right).$$

V-statistic

- Estimator for $\mathbb{E}_{\mathbb{P}} h(X_1, \ldots, X_m)$:

$$V_n = V(x_1, \ldots, x_n) = \frac{1}{n^m} \sum_{i_1=1}^{m} \ldots \sum_{i_m=1}^{n} h\left(x_{i_1}, \ldots, x_{i_m}\right).$$

- Samples with replacement.

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int(x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2X$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2X = \frac{\mathbb{E}X_1^2 + \mathbb{E}X_2^2}{2} - \mathbb{E}X_1\mathbb{E}X_2$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample average:

$$h(x) = x, \qquad U(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2 X = \frac{\mathbb{E}X_1^2 + \mathbb{E}X_2^2}{2} - \mathbb{E}X_1\mathbb{E}X_2 = \mathbb{E}h(X_1, X_2),$$

$$h(x_1, x_2) = \frac{x_1^2 + x_2^2 - 2x_1x_2}{2}$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2X = \frac{\mathbb{E}X_1^2 + \mathbb{E}X_2^2}{2} - \mathbb{E}X_1\mathbb{E}X_2 = \mathbb{E}h(X_1,X_2),$$

$$h(x_1,x_2) = \frac{x_1^2 + x_2^2 - 2x_1x_2}{2} = \frac{(x_1-x_2)^2}{2},$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2 X = \frac{\mathbb{E}X_1^2 + \mathbb{E}X_2^2}{2} - \mathbb{E}X_1\mathbb{E}X_2 = \mathbb{E}h(X_1,X_2),$$

$$h(x_1,x_2) = \frac{x_1^2 + x_2^2 - 2x_1x_2}{2} = \frac{(x_1-x_2)^2}{2},$$

$$U(x_1,\ldots,x_n) = \frac{2}{n(n-1)} \sum_{1\leqslant i<j\leqslant n} h(x_i,x_j)$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}} X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}} X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2 X = \frac{\mathbb{E}X_1^2 + \mathbb{E}X_2^2}{2} - \mathbb{E}X_1\mathbb{E}X_2 = \mathbb{E}h(X_1,X_2),$$

$$h(x_1,x_2) = \frac{x_1^2 + x_2^2 - 2x_1x_2}{2} = \frac{(x_1-x_2)^2}{2},$$

$$U(x_1,\ldots,x_n) = \frac{2}{n(n-1)}\sum_{1\leqslant i<j\leqslant n} h(x_i,x_j) = \frac{1}{n(n-1)}\sum_{i\neq j} h(x_i,x_j)$$

U-statistic: examples

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample average:

$$h(x) = x, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- $\theta(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}X^k$. Sample k^{th} moment:

$$h(x) = x^k, \qquad U(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- $\theta(\mathbb{P}) = \sigma^2(\mathbb{P}) = \int (x-\mu)^2 \mathrm{d}\mathbb{P}(x)$, $\mu = \mathbb{E}_{X\sim\mathbb{P}}X$. Sample variance:

$$\sigma^2(\mathbb{P}) = \mathbb{E}X^2 - \mathbb{E}^2X = \frac{\mathbb{E}X_1^2 + \mathbb{E}X_2^2}{2} - \mathbb{E}X_1\mathbb{E}X_2 = \mathbb{E}h(X_1,X_2),$$

$$h(x_1,x_2) = \frac{x_1^2 + x_2^2 - 2x_1x_2}{2} = \frac{(x_1-x_2)^2}{2},$$

$$U(x_1,\ldots,x_n) = \frac{2}{n(n-1)}\sum_{1\leqslant i<j\leqslant n} h(x_i,x_j) = \frac{1}{n(n-1)}\sum_{i\neq j} h(x_i,x_j) = s_n^2.$$

U-statistic: examples+

$$\theta(\mathbb{P}) = F_{\mathbb{P}}(t_0) = \mathbb{P}(X \leqslant t_0)$$

U-statistic: examples+

$$\theta(\mathbb{P}) = F_{\mathbb{P}}(t_0) = \mathbb{P}(X \leqslant t_0) = \mathbb{E}_{\mathbb{P}}\chi_{(-\infty,t_0]},$$

U-statistic: examples+

$$\theta(\mathbb{P}) = F_{\mathbb{P}}(t_0) = \mathbb{P}(X \leqslant t_0) = \mathbb{E}_{\mathbb{P}} \chi_{(-\infty, t_0]},$$

$$h(x) = \chi_{\{x \leqslant t_0\}},$$

$$U(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \chi_{\{x_i \leqslant t_0\}}$$

U-statistic: examples+

$$\theta(\mathbb{P}) = F_{\mathbb{P}}(t_0) = \mathbb{P}(X \leqslant t_0) = \mathbb{E}_{\mathbb{P}} \chi_{(-\infty, t_0]},$$

$$h(x) = \chi_{\{x \leqslant t_0\}},$$

$$U(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \chi_{\{x_i \leqslant t_0\}} = F_n(t_0),$$

F_n: empirical cdf.

Extension: if we have L independent samples

- Given: $x_1^{(j)}, \ldots, x_{m_j}^{(j)} \stackrel{i.i.d.}{\sim} \mathbb{P}_j$ $(j = 1, \ldots, L)$, $n_i \geqslant m_i$.

Extension: if we have L independent samples

- Given: $x_1^{(j)}, \ldots, x_{m_j}^{(j)} \overset{i.i.d.}{\sim} \mathbb{P}_j$ $(j = 1, \ldots, L)$, $n_i \geqslant m_i$.
- Goal: estimate: $\theta = \mathbb{E}h\left(X_1^{(1)}, \ldots, X_{m_1}^{(1)}, \ldots, X_1^{(L)}, \ldots, X_{m_L}^{(L)}\right)$.

Extension: if we have L independent samples

- Given: $x_1^{(j)}, \ldots, x_{m_j}^{(j)} \stackrel{i.i.d.}{\sim} \mathbb{P}_j$ $(j = 1, \ldots, L)$, $n_i \geqslant m_i$.
- Goal: estimate: $\theta = \mathbb{E}h\left(X_1^{(1)}, \ldots, X_{m_1}^{(1)}, \ldots, X_1^{(L)}, \ldots, X_{m_L}^{(L)}\right)$.
- Assumption: symmetry for each block.

Extension: if we have L independent samples

- Given: $x_1^{(j)}, \ldots, x_{m_j}^{(j)} \overset{i.i.d.}{\sim} \mathbb{P}_j$ $(j = 1, \ldots, L)$, $n_i \geqslant m_i$.
- Goal: estimate: $\theta = \mathbb{E}h\left(X_1^{(1)}, \ldots, X_{m_1}^{(1)}, \ldots, X_1^{(L)}, \ldots, X_{m_L}^{(L)}\right)$.
- Assumption: symmetry for each block.
- L-sample U-statistic

$$U_n = \frac{1}{\prod_{j=1}^{L} \binom{n_j}{m_j}} \sum_{c} h\left(X_1^{(1)}, \ldots, X_{m_1}^{(1)}, \ldots, X_1^{(L)}, \ldots, X_{m_L}^{(L)}\right).$$

Results [Serfling, 1980]

U_n: minimum variance unbiased estimator.

Results [Serfling, 1980]

U_n: minimum variance unbiased estimator. Notation:

- Asymptotics: depends on $\mathrm{var} \neq 0$ condition.

$$h_c(x_1, \ldots, x_c) := \mathbb{E}h\left(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m\right),$$

Results [Serfling, 1980]

U_n: minimum variance unbiased estimator. Notation:

- Asymptotics: depends on $\mathrm{var} \neq 0$ condition.

$$h_c(x_1,\ldots,x_c) := \mathbb{E}h\left(x_1,\ldots,x_c,X_{c+1},\ldots,X_m\right),$$
$$v_c := \mathrm{var}\, h_c\left(X_1,\ldots,X_c\right),\; v_0 = 0.$$

Results [Serfling, 1980]

U_n: minimum variance unbiased estimator. Notation:

- Asymptotics: depends on $\mathrm{var} \neq 0$ condition.

$$h_c(x_1, \ldots, x_c) := \mathbb{E}h\left(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m\right),$$
$$v_c := \mathrm{var}\, h_c\left(X_1, \ldots, X_c\right), v_0 = 0.$$

- If $\mathbb{E}h^2(X_1, \ldots, X_m) < \infty$:

$$0 = v_0 \leqslant v_1 \leqslant \ldots \leqslant v_m = \mathrm{var}\, h\left(X_1, \ldots, X_m\right) < \infty.$$

Results [Serfling, 1980]

U_n: minimum variance unbiased estimator. Notation:

- Asymptotics: depends on $\mathrm{var} \neq 0$ condition.

$$h_c(x_1,\ldots,x_c) := \mathbb{E}h\left(x_1,\ldots,x_c,X_{c+1},\ldots,X_m\right),$$
$$v_c := \mathrm{var}\, h_c\left(X_1,\ldots,X_c\right), v_0 = 0.$$

- If $\mathbb{E}h^2(X_1,\ldots,X_m) < \infty$:

$$0 = v_0 \leqslant v_1 \leqslant \ldots \leqslant v_m = \mathrm{var}\, h\left(X_1,\ldots,X_m\right) < \infty.$$

- c: $0 = v_1 = \ldots = v_{c-1} < v_c$. $c = 1$: non-degenerate, $c \geqslant 2$: degenerate U-statistic.

Results [Serfling, 1980]

U_n: minimum variance unbiased estimator. Notation:

- Asymptotics: depends on $\mathrm{var} \neq 0$ condition.

$$h_c(x_1, \ldots, x_c) := \mathbb{E}h\left(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m\right),$$
$$v_c := \mathrm{var}\, h_c\left(X_1, \ldots, X_c\right), v_0 = 0.$$

- If $\mathbb{E}h^2(X_1, \ldots, X_m) < \infty$:

$$0 = v_0 \leqslant v_1 \leqslant \ldots \leqslant v_m = \mathrm{var}\, h\left(X_1, \ldots, X_m\right) < \infty.$$

- c: $0 = v_1 = \ldots = v_{c-1} < v_c$. $c = 1$: non-degenerate, $c \geqslant 2$: degenerate U-statistic.

In most applications

$c = 1$ or $c = 2$.

Asymptotics for $c = 1$

Assume: $\mathbb{E}_{\mathbb{P}} h^2 < \infty$, $c = 1$.

$$n^{\frac{1}{2}} (U_n - \theta) \xrightarrow{d} N\left(0, m^2 v_1\right),$$

i.e.

$$U_n \text{ is } AN\left(\theta, \frac{m^2 v_1}{n}\right),$$

AN := asymptotically normal.

Asymptotics for $c = 2$

Assume: $\mathbb{E}_{\mathbb{P}} h^2 < \infty$, $c = 2$.

$$n(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2} Y, \qquad Y = \sum_{j=1}^{\infty} \lambda_j (\chi_j^2 - 1),$$

where

Asymptotics for $c = 2$

Assume: $\mathbb{E}_{\mathbb{P}} h^2 < \infty$, $c = 2$.

$$n(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2} Y, \qquad Y = \sum_{j=1}^{\infty} \lambda_j (\chi_j^2 - 1),$$

where

- χ_j^2: i.i.d. $N^2(0,1)$ variables,

Asymptotics for $c = 2$

Assume: $\mathbb{E}_{\mathbb{P}} h^2 < \infty$, $c = 2$.

$$n(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2} Y, \qquad Y = \sum_{j=1}^{\infty} \lambda_j (\chi_j^2 - 1),$$

where

- χ_j^2: i.i.d. $N^2(0,1)$ variables,
- λ_j: $\mathbb{R}$-eigenvalues of $T = T\left(\tilde{h}_2\right)$, $\tilde{h}_2 = h_2 - \theta$

$$(Tg)(x) = \int \tilde{h}_2(x,y) g(y) \mathrm{d}\mathbb{P}(y), \quad g \in L^2.$$

Exponential bound for U-statistic

Theorem (Hoeffding inequality)

Let $h(x_1, \ldots, x_m) \in [a, b]$. *If* $\sigma^2 = \operatorname{var} h$, *then for any* $t > 0$

$$\mathbb{P}(U_n - \theta \geqslant t) \leqslant e^{-\frac{2[n/m]t^2}{(b-a)^2}}.$$

U-statistic: local summary

- Minimum variance unbiased estimator.
- $c = 1$: asymptotically normal.
- $c = 2$: asymptotically ∞-sum of weighted χ^2.
- For bounded h: Hoeffding inequality.

Application

Hypothesis testing!

Hypothesis testing

What is a two-sample test?

- Given:
  - $X = \{x_i\}_{i=1}^m \stackrel{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \stackrel{i.i.d.}{\sim} \mathbb{Q}$.
  - Example: $x_i = i^{th}$ happy face, $y_j = j^{th}$ sad face.

What is a two-sample test?

- Given:
  - $X = \{x_i\}_{i=1}^m \stackrel{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \stackrel{i.i.d.}{\sim} \mathbb{Q}$.
  - Example: $x_i = i^{th}$ happy face, $y_j = j^{th}$ sad face.
- Problem: using X, Y test

$$H_0 : \mathbb{P} = \mathbb{Q}, \text{ vs}$$
$$H_1 : \mathbb{P} \neq \mathbb{Q}.$$

What is a two-sample test?

- Given:
  - $X = \{x_i\}_{i=1}^m \stackrel{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \stackrel{i.i.d.}{\sim} \mathbb{Q}$.
  - Example: $x_i = i^{th}$ happy face, $y_j = j^{th}$ sad face.
- Problem: using X, Y test

$$H_0 : \mathbb{P} = \mathbb{Q}, \text{ vs}$$
$$H_1 : \mathbb{P} \neq \mathbb{Q}.$$

- Assumption: $x, y \in \mathcal{X}$. $(\mathcal{X}, k)$: kernel-endowed domain.

What is a two-sample test?

- Given:
  - $X = \{x_i\}_{i=1}^m \overset{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}$.
  - Example: $x_i = i^{th}$ happy face, $y_j = j^{th}$ sad face.
- Problem: using X, Y test

$$H_0 : \mathbb{P} = \mathbb{Q}, \text{ vs}$$
$$H_1 : \mathbb{P} \neq \mathbb{Q}.$$

- Assumption: $x, y \in \mathcal{X}$. $(\mathcal{X}, k)$: kernel-endowed domain.

Discrepancy measure

Example: MMD

What is an independence test?

- Given: paired samples
  - $Z = \{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}_{xy}$.
  - Example:
    - x_i: i^{th} text in English, y_i: i^{th} text translated to French.

What is an independence test?

- Given: paired samples
  - $Z = \{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}_{xy}$.
  - Example:
    - x_i: i^{th} text in English, y_i: i^{th} text translated to French.
- Problem: using data Z test

$$H_0 : \mathbb{P}_{xy} = \mathbb{P}_x \otimes \mathbb{P}_y, \qquad H_1 : \mathbb{P}_{xy} \neq \mathbb{P}_x \otimes \mathbb{P}_y.$$

What is an independence test?

- Given: paired samples
  - $Z = \{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}_{xy}$.
  - Example:
    - x_i: i^{th} text in English, y_i: i^{th} text translated to French.
- Problem: using data Z test

$$H_0 : \mathbb{P}_{xy} = \mathbb{P}_x \otimes \mathbb{P}_y, \qquad H_1 : \mathbb{P}_{xy} \neq \mathbb{P}_x \otimes \mathbb{P}_y.$$

- Assumption: $(x, y) \in \mathcal{X} \times \mathcal{Y}$. $(\mathcal{X}, k)$, $(\mathcal{Y}, \ell)$: with kernels.

What is an independence test?

- Given: paired samples
  - $Z = \{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}_{xy}$.
  - Example:
    - x_i: i^{th} text in English, y_i: i^{th} text translated to French.
- Problem: using data Z test

$$H_0 : \mathbb{P}_{xy} = \mathbb{P}_x \otimes \mathbb{P}_y, \qquad H_1 : \mathbb{P}_{xy} \neq \mathbb{P}_x \otimes \mathbb{P}_y.$$

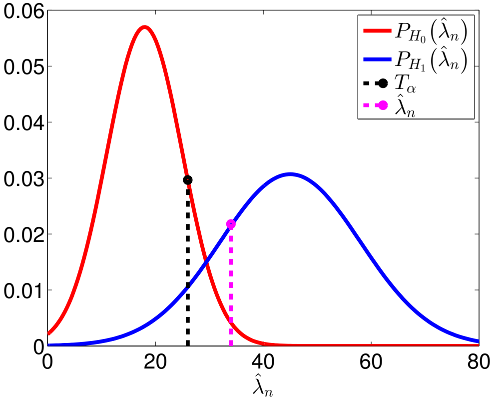
- Assumption: $(x, y) \in \mathcal{X} \times \mathcal{Y}$. $(\mathcal{X}, k)$, $(\mathcal{Y}, \ell)$: with kernels.

Discrepancy measure

Example: HSIC

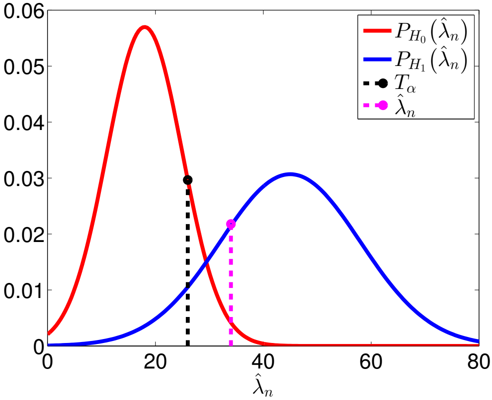
Concepts in hypothesis testing

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- Under H_0: $P_{H_0}\big(\underbrace{\hat{\lambda}_n \leq T_\alpha}_{\text{correctly accepting } H_0}\big) = 1 - \alpha$.



Concepts in hypothesis testing

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- Under H_0: $P_{H_0}\Big(\underbrace{\hat{\lambda}_n \leqslant T_\alpha}_{\text{correctly accepting } H_0}\Big) = 1 - \alpha$.
- Under H_1: $P_{H_1}(T_\alpha < \hat{\lambda}_n) = P(\text{correctly rejecting } H_0) =: \text{power}$.



Two-sample testing (aka homogeneity testing) – details.

Two-sample testing with MMD [Gretton et al., 2007, Gretton et al., 2012]

- Statistic: $\lambda_n = \widehat{MMD}_b^2$ or $\widehat{MMD}_u^2$.

Two-sample testing with MMD [Gretton et al., 2007, Gretton et al., 2012]

- Statistic: $\lambda_n = \widehat{MMD}_b^2$ or $\widehat{MMD}_u^2$.
- Reject H_0: if λ_n is 'large'.

Two-sample testing with MMD [Gretton et al., 2007, Gretton et al., 2012]

- Statistic: $\lambda_n = \widehat{MMD}_b^2$ or $\widehat{MMD}_u^2$.
- Reject H_0: if λ_n is 'large'.
- We need to control λ_n.

Two-sample testing with MMD [Gretton et al., 2007, Gretton et al., 2012]

- Statistic: $\lambda_n = \widehat{MMD_b^2}$ or $\widehat{MMD_u^2}$.
- Reject H_0: if λ_n is 'large'.
- We need to control λ_n.
- We will use U-statistic theory.

Finite-sample control

- Large deviation inequalities.
- $P\left(\left|\widehat{MMD}(\mathbb{P},\mathbb{Q}) - MMD(\mathbb{P},\mathbb{Q})\right| \geqslant \epsilon\right) \leqslant f(\epsilon, m, n) \xrightarrow{m,n\to\infty} 0.$

Finite-sample control

- Large deviation inequalities.
- $P\left(\left|\widehat{MMD}(\mathbb{P},\mathbb{Q}) - MMD(\mathbb{P},\mathbb{Q})\right| \geqslant \epsilon\right) \leqslant f(\epsilon, m, n) \xrightarrow{m,n\to\infty} 0.$
- $\Rightarrow$ tests: consistent against fixed alternative.

Finite-sample control

- Distribution-free tests
  - quantile: $\mathbb{P}, \mathbb{Q}$-independent!
  - less sensitive to detecting differences.

Finite-sample control

- Distribution-free tests
  - quantile: $\mathbb{P}, \mathbb{Q}$-independent!
  - less sensitive to detecting differences.
- Proof idea:

Finite-sample control

- Distribution-free tests
  - quantile: $\mathbb{P}, \mathbb{Q}$-independent!
  - less sensitive to detecting differences.
- Proof idea:
  - $\widehat{MMD}_b^2$: bounded difference property, McDiarmid inequality.

Finite-sample control

- Distribution-free tests
  - quantile: $\mathbb{P}, \mathbb{Q}$-independent!
  - less sensitive to detecting differences.
- Proof idea:
  - $\widehat{MMD}_b^2$: bounded difference property, McDiarmid inequality.
  - $\widehat{MMD}_u^2$: large deviation bound of U-statistics.

Asymptotics based test

Needed: Asymptotic distribution of $\widehat{MMD}_u^2$.

$$\widehat{MMD}_u^2(\mathbb{P},\mathbb{Q}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j).$$

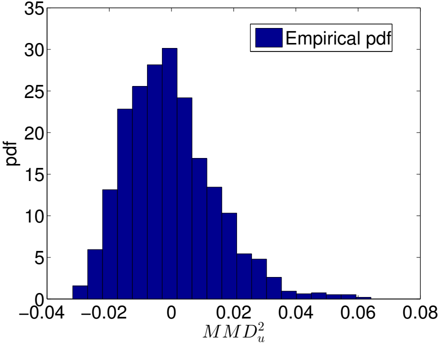
Two-sample test using MMD asymptotics: H_0 [$c = 2$!]

Under H_0 ($\mathbb{P} = \mathbb{Q}$): asymptotic distribution is

$$n\widehat{MMD}_u^2(\mathbb{P}, \mathbb{P}) \sim \sum_{i=1}^{\infty} \lambda_i (z_i^2 - 2),$$

where $z_i \sim N(0, 2)$ i.i.d.,

$$\int_{\mathcal{X}} \tilde{k}(x, x') v_i(x) \mathrm{d}\mathbb{P}(x) = \lambda_i v_i(x'), \; \tilde{k}(x, x') = \langle \varphi_x - \mu_{\mathbb{P}}, \varphi_{x'} - \mu_{\mathbb{P}} \rangle_{\mathcal{H}_k}.$$



In practice

Approximate the null by

- permutation-test: slow.

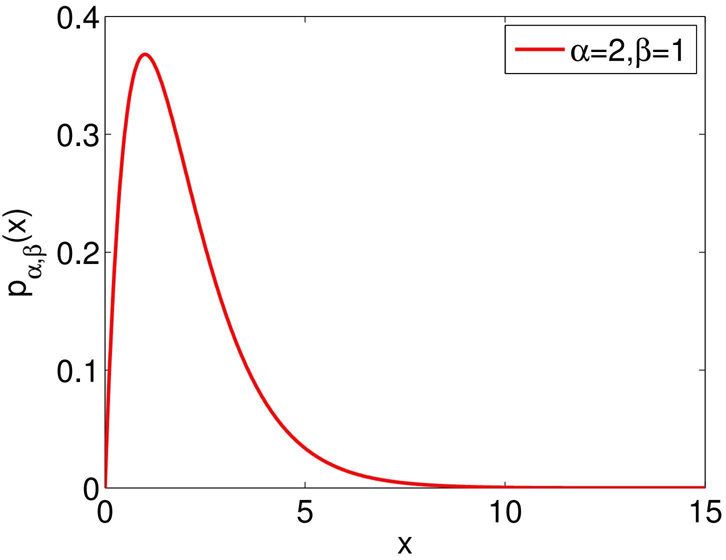
In practice

Approximate the null by

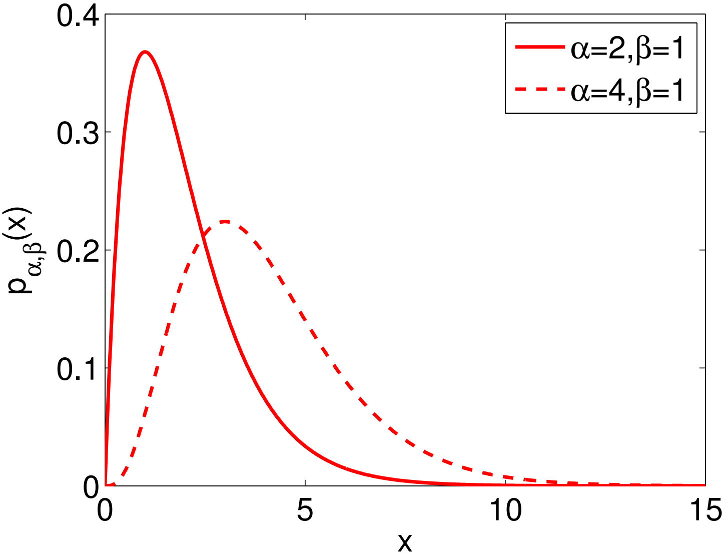
- permutation-test: slow.
- two-parameter gamma distribution [Johnson et al., 1994]:

$$p_{\alpha,\beta}(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \quad (x > 0, \alpha\text{: shape} > 0, \beta\text{: scale} > 0).$$

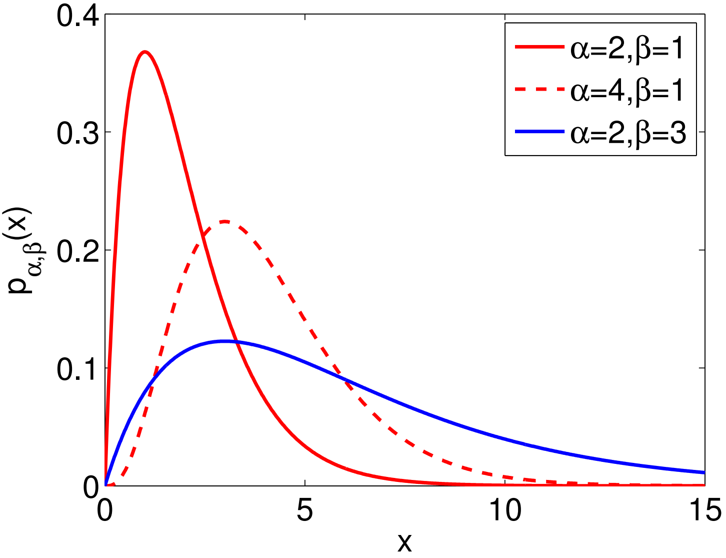
Gamma distribution: demo



Gamma distribution: demo



Gamma distribution: demo



Gamma approximation

- Assumption: statistic $T \sim p_{\alpha,\beta}$.

Gamma approximation

- Assumption: statistic $T \sim p_{\alpha,\beta}$.
- For $p_{\alpha,\beta}$ gamma distribution:

$$\mathbb{E}T = \alpha\beta, \quad \mathrm{var}(T) = \alpha\beta^2$$

Gamma approximation

- Assumption: statistic $T \sim p_{\alpha,\beta}$.
- For $p_{\alpha,\beta}$ gamma distribution:

$$\mathbb{E}T = \alpha\beta, \quad \mathrm{var}(T) = \alpha\beta^2 \quad \Rightarrow \quad \beta = \frac{\mathrm{var}(T)}{\mathbb{E}T} \quad \Rightarrow \quad \alpha = \frac{\mathbb{E}^2 T}{\mathrm{var}(T)}.$$

Gamma approximation

- Assumption: statistic $T \sim p_{\alpha,\beta}$.
- For $p_{\alpha,\beta}$ gamma distribution:

$$\mathbb{E}T = \alpha\beta, \quad \mathrm{var}(T) = \alpha\beta^2 \quad \Rightarrow \quad \beta = \frac{\mathrm{var}(T)}{\mathbb{E}T} \quad \Rightarrow \quad \alpha = \frac{\mathbb{E}^2 T}{\mathrm{var}(T)}.$$

- Thus, $\widehat{\mathbb{E}T}$ and $\widehat{\mathrm{var}(T)} \rightarrow \hat{\alpha}, \hat{\beta}$.

Gamma approximation

- Assumption: statistic $T \sim p_{\alpha,\beta}$.
- For $p_{\alpha,\beta}$ gamma distribution:

$$\mathbb{E}T = \alpha\beta, \quad \mathrm{var}(T) = \alpha\beta^2 \quad \Rightarrow \quad \beta = \frac{\mathrm{var}(T)}{\mathbb{E}T} \quad \Rightarrow \quad \alpha = \frac{\mathbb{E}^2 T}{\mathrm{var}(T)}.$$

- Thus, $\widehat{\mathbb{E}T}$ and $\widehat{\mathrm{var}(T)} \rightarrow \hat{\alpha}, \hat{\beta}$.
- Consistency of the test is lost.

Which null approximation to use?

Rules-of-thumb:

- Small sample size: permutation test.

Which null approximation to use?

Rules-of-thumb:

- Small sample size: permutation test.
- Medium sample size: gamma approximation, truncated expansion [Gretton et al., 2009],

Which null approximation to use?

Rules-of-thumb:

- Small sample size: permutation test.
- Medium sample size: gamma approximation, truncated expansion [Gretton et al., 2009],
- Large sample size:
  - online techniques [Gretton et al., 2012], or
  - recent linear methods (next time).

Independence testing: HSIC

Independence testing

Theorem ([Gretton et al., 2008, Pfister et al., 2017])

Under H_0

$$n\widehat{HSIC}_b^2 \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i z_i^2, \quad z_i \sim N(0,1).$$

Independence testing

Theorem ([Gretton et al., 2008, Pfister et al., 2017])

Under H_0

$$n\widehat{HSIC}_b^2 \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i z_i^2, \quad z_i \sim N(0,1).$$

Notes:

- For U-statistic: $\sum_i \lambda_i (z_i^2 - 1)$.

Independence testing

Theorem ([Gretton et al., 2008, Pfister et al., 2017])

Under H_0

$$n\widehat{HSIC}_b^2 \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i z_i^2, \quad z_i \sim N(0,1).$$

Notes:

- For U-statistic: $\sum_i \lambda_i (z_i^2 - 1)$.
- In practice: permutation-test/gamma-approximation.

Related work

Two-sample problem: truncated expansion

[Gretton et al., 2009]: $n = m$, $z_i = (x_i, y_i)$. Estimator:

$$\widehat{MMD}^2_{u'}(\mathbb{P},\mathbb{Q}) = \frac{1}{n(n-1)} \sum_{i \neq j} h(z_i, z_j),$$

$$h(z, z') = k(x, x') + k(y, y') - k(x, y') - k(x', y).$$

Two-sample problem: truncated expansion

[Gretton et al., 2009]: $n = m$, $z_i = (x_i, y_i)$. Estimator:

$$\widehat{MMD^2_{u'}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{n(n-1)} \sum_{i \neq j} h(z_i, z_j),$$

$$h(z, z') = k(x, x') + k(y, y') - k(x, y') - k(x', y).$$

$\widehat{MMD^2_{u'}}$: unbiased.

Two-sample problem: truncated expansion – continued

Theorem

Assuming $\sum_{i=1}^{\infty} \lambda_i^{\frac{1}{2}} < \infty$, *the empirical null converges as* $n \to \infty$

$$T_n := \sum_{i=1}^{n} \hat{\lambda}_{i,n} \left(a_i^2 - 2\right) \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i \left(a_i^2 - 2\right), \quad a_i \sim N(0,2).$$

Two-sample problem: truncated expansion – continued

Theorem

Assuming $\sum_{i=1}^{\infty} \lambda_i^{\frac{1}{2}} < \infty$, *the empirical null converges as* $n \to \infty$

$$T_n := \sum_{i=1}^{n} \hat{\lambda}_{i,n} \left(a_i^2 - 2\right) \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i \left(a_i^2 - 2\right), \quad a_i \sim N(0,2).$$

Note:

$$\hat{\lambda}_{i,n} := \frac{\lambda_i(\tilde{\mathbf{G}}_x)}{n} \quad (i = 1, \ldots, n), \quad \tilde{\mathbf{G}}_x \in \mathbb{R}^{n \times n}.$$

Online variant [Gretton et al., 2012]

$$\widehat{MMD}^2_{u'}(\mathbb{P}, \mathbb{Q}) = \frac{1}{n(n-1)} \sum_{i \neq j} h(z_i, z_j),$$

has a natural online approximation, $n_2 := [n/2]$

$$\widehat{MMD}^2_l(\mathbb{P}, \mathbb{Q}) = \frac{1}{n_2} \sum_{i=1}^{n_2} h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i})).$$

Online variant [Gretton et al., 2012]

$$\widehat{MMD}^2_{u'}(\mathbb{P},\mathbb{Q}) = \frac{1}{n(n-1)}\sum_{i\neq j} h(z_i,z_j),$$

has a natural online approximation, $n_2 := [n/2]$

$$\widehat{MMD}^2_l(\mathbb{P},\mathbb{Q}) = \frac{1}{n_2}\sum_{i=1}^{n_2} h((x_{2i-1},y_{2i-1}),(x_{2i},y_{2i})).$$

- Unbiased.

Online variant [Gretton et al., 2012]

$$\widehat{MMD}^2_{u'}(\mathbb{P},\mathbb{Q}) = \frac{1}{n(n-1)}\sum_{i\neq j} h(z_i, z_j),$$

has a natural online approximation, $n_2 := [n/2]$

$$\widehat{MMD}^2_{l}(\mathbb{P},\mathbb{Q}) = \frac{1}{n_2}\sum_{i=1}^{n_2} h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i})).$$

- Unbiased.
- Linear-time: streaming data.

Online variant [Gretton et al., 2012]

$$\widehat{MMD}^2_{u'}(\mathbb{P},\mathbb{Q}) = \frac{1}{n(n-1)}\sum_{i\neq j} h(z_i, z_j),$$

has a natural online approximation, $n_2 := [n/2]$

$$\widehat{MMD}^2_{l}(\mathbb{P},\mathbb{Q}) = \frac{1}{n_2}\sum_{i=1}^{n_2} h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i})).$$

- Unbiased.
- Linear-time: streaming data.
- In practice: high variance.

Online variant – continued

By the average the CLT kicks in:

Theorem

Assuming $\mathbb{E}h^2 \in (0, \infty)$, $\widehat{MMD}_l^2$ *is asymptotically normal* (H_0/H_1)

$$\sqrt{n}\left[\widehat{MMD}_l^2(\mathbb{P},\mathbb{Q}) - MMD^2(\mathbb{P},\mathbb{Q})\right] \xrightarrow{d} N\left(0,\sigma^2\right),$$

where $\sigma^2 = 2\left[\mathbb{E}_{z,z'}h^2\left(z,z'\right) - \mathbb{E}_{z,z'}^2 h\left(z,z'\right)\right]$.

Block version [Zaremba et al., 2013]

Idea:

- partition the data to blocks of size B,
- on each block: compute $\widehat{MMD}_l^2$,
- average the results.

Block version – continued

Properties:

- Statistic: asymptotically normal (H_0, H_1).
- For consistency: increase B_m s.t. $\frac{m}{B_m} \to \infty$.
- Reduced variance.

Three-variable interaction test

- Goal (interaction):

$$([x_1; x_2] \perp x_3) \vee ([x_1; x_3] \perp x_2) \vee ([x_2; x_3] \perp x_1).$$

Example: $\mathbb{P} = \mathbb{P}_{12} \otimes \mathbb{P}_3$.

Three-variable interaction test

- Goal (interaction):

$$([x_1; x_2] \perp x_3) \vee ([x_1; x_3] \perp x_2) \vee ([x_2; x_3] \perp x_1).$$

  Example: $\mathbb{P} = \mathbb{P}_{12} \otimes \mathbb{P}_3$.
- Applications:
  - structure learning of graphical models,
  - discovering V-structures.

Three-variable interaction test – continued

Analogy

Independence $\Leftrightarrow \mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \Leftrightarrow \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2 = 0.$

Three-variable interaction test – continued

Analogy

Independence $\Leftrightarrow \mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \Leftrightarrow \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2 = 0$.

- Lancaster 3-variable interaction [Lancaster, 1969]:

$$L(\mathbb{P}) = \mathbb{P} - \mathbb{P}_{1,2} \otimes \mathbb{P}_3 - \mathbb{P}_{2,3} \otimes \mathbb{P}_1 - \mathbb{P}_{1,3} \otimes \mathbb{P}_2 + 2\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3.$$

is a signed measure,

$$\text{interaction} \Rightarrow L(\mathbb{P}) = 0.$$

Three-variable interaction test – continued

Analogy

Independence $\Leftrightarrow \mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \Leftrightarrow \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2 = 0$.

- Lancaster 3-variable interaction [Lancaster, 1969]:

  $$L(\mathbb{P}) = \mathbb{P} - \mathbb{P}_{1,2} \otimes \mathbb{P}_3 - \mathbb{P}_{2,3} \otimes \mathbb{P}_1 - \mathbb{P}_{1,3} \otimes \mathbb{P}_2 + 2\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3.$$

  is a signed measure,

  $$\text{interaction} \Rightarrow L(\mathbb{P}) = 0.$$

- $x_i \in (\mathcal{X}_i, k_i)$ are kernel endowed domains.

Three-variable interaction test – continued

- Interaction index [Sejdinovic et al., 2013a]:

$$I = \left\| \mu_{L(\mathbb{P})} \right\|^2_{\mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2} \otimes \mathcal{H}_{k_3}}.$$

Three-variable interaction test – continued

- Interaction index [Sejdinovic et al., 2013a]:

$$I = \left\| \mu_{L(\mathbb{P})} \right\|^2_{\mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2} \otimes \mathcal{H}_{k_3}}.$$

- Empirical estimate:

$$\hat{I} = \frac{\left(\tilde{\mathbf{G}}_{x_1} \circ \tilde{\mathbf{G}}_{x_3} \circ \tilde{\mathbf{G}}_{x_3} \right)_{++}}{n^2}.$$

Three-variable interaction test – continued

- Interaction index [Sejdinovic et al., 2013a]:

$$I = \left\| \mu_{L(\mathbb{P})} \right\|^2_{\mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2} \otimes \mathcal{H}_{k_3}}.$$

- Empirical estimate:

$$\hat{I} = \frac{\left(\tilde{\mathbf{G}}_{x_1} \circ \tilde{\mathbf{G}}_{x_3} \circ \tilde{\mathbf{G}}_{x_3} \right)_{++}}{n^2}.$$

- Null approximation: permutation-test.

Time-series tests: independence

- Goal: test the stationary distribution of processes.
- Independence tests
  - Statistic: HSIC.

Time-series tests: independence

- Goal: test the stationary distribution of processes.
- Independence tests
  - Statistic: HSIC.
  - i.i.d. permutation technique: would fail.

Time-series tests: independence

- Goal: test the stationary distribution of processes.
- Independence tests
  - Statistic: HSIC.
  - i.i.d. permutation technique: would fail.
  - Idea: shift-approach = preserve 'time structure' [Chwialkowski and Gretton, 2014].

Time-series tests: two-sample, independence, interaction

- Permutation approach (i.i.d): ± 1.

Time-series tests: two-sample, independence, interaction

- Permutation approach (i.i.d): ± 1.
- Idea: mask according to the memory of the processes.

Time-series tests: two-sample, independence, interaction

- Permutation approach (i.i.d): ± 1.
- Idea: mask according to the memory of the processes.
- Implementation [Chwialkowski et al., 2014]: based on wild bootstrap [Leucht and H.Neumann, 2013].

Time-series tests: two-sample, independence, interaction

- Permutation approach (i.i.d): ± 1.
- Idea: mask according to the memory of the processes.
- Implementation [Chwialkowski et al., 2014]: based on wild bootstrap [Leucht and H.Neumann, 2013].

3-variable interaction:

- Lancaster interaction + wild bootstrap [Rubenstein et al., 2016].

Goodness-of-fit test

- Given:
  - $\{x_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} q$,
  - p: target distribution.

Goodness-of-fit test

- Given:
  - $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$,
  - p: target distribution.
- p, q live on $\mathcal{X} \subset \mathbb{R}^d$ (differentiability), kernel k on $\mathcal{X}$.

Goodness-of-fit test

- Given:
  - $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$,
  - p: target distribution.
- p, q live on $\mathcal{X} \subset \mathbb{R}^d$ (differentiability), kernel k on $\mathcal{X}$.
- Goal:

$$H_0 : p = q,$$
$$H_1 : p \neq q.$$

Goodness-of-fit test: continued

- Idea [Chwialkowski et al., 2016, Liu et al., 2016]: Stein operator

$$(S_p f)(x) = \sum_{i=1}^{d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right], \quad f \in \mathcal{H} := \otimes_{i=1}^{d} \mathcal{H}_k,$$

Goodness-of-fit test: continued

- Idea [Chwialkowski et al., 2016, Liu et al., 2016]: Stein operator

$$(S_p f)(x) = \sum_{i=1}^{d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right], \quad f \in \mathcal{H} := \otimes_{i=1}^{d} \mathcal{H}_k,$$

$$\mathbb{E}_{x \sim q} (S_p f)(x) = 0 \Leftarrow p = q \, (\forall f).$$

Goodness-of-fit test: continued

- Idea [Chwialkowski et al., 2016, Liu et al., 2016]: Stein operator

$$(S_p f)(x) = \sum_{i=1}^{d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right], \quad f \in \mathcal{H} := \otimes_{i=1}^{d} \mathcal{H}_k,$$

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}} \leqslant 1} \mathbb{E}_{x \sim q} (S_p f)(x).$$

Goodness-of-fit test: continued

- Idea [Chwialkowski et al., 2016, Liu et al., 2016]: Stein operator

$$(S_p f)(x) = \sum_{i=1}^{d}\left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i}\right], \quad f \in \mathcal{H} := \otimes_{i=1}^{d}\mathcal{H}_k,$$

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}} \leqslant 1} \mathbb{E}_{x\sim q}(S_p f)(x).$$

- For c_0-universal k: $T_p(q) = 0 \Leftrightarrow p = q$.

Goodness-of-fit test: continued

- Idea [Chwialkowski et al., 2016, Liu et al., 2016]: Stein operator

$$(S_p f)(x) = \sum_{i=1}^{d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right], \quad f \in \mathcal{H} := \otimes_{i=1}^{d} \mathcal{H}_k,$$

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}} \leqslant 1} \mathbb{E}_{x \sim q} (S_p f)(x).$$

- For c_0-universal k: $T_p(q) = 0 \Leftrightarrow p = q$.
- Enough: p up to multiplicative constant ($\nabla \log p$).

Goodness-of-fit test: continued

- Idea [Chwialkowski et al., 2016, Liu et al., 2016]: Stein operator

$$(S_p f)(x) = \sum_{i=1}^{d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right], \quad f \in \mathcal{H} := \otimes_{i=1}^{d} \mathcal{H}_k,$$

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}} \leqslant 1} \mathbb{E}_{x \sim q} (S_p f)(x).$$

- For c_0-universal k: $T_p(q) = 0 \Leftrightarrow p = q$.
- Enough: p up to multiplicative constant ($\nabla \log p$).
- Null approximation: wild bootstrap (including non-i.i.d.).

Quadratic-time methods

- Two-sample, independence, interaction, goodness-of-fit test.

Quadratic-time methods

- Two-sample, independence, interaction, goodness-of-fit test.
- Kernel endowed domain (goodness-of-fit $\mathbb{R}^d$).

Quadratic-time methods

- Two-sample, independence, interaction, goodness-of-fit test.
- Kernel endowed domain (goodness-of-fit $\mathbb{R}^d$).
- Typically: null can be ugly, techniques do not scale well.

Quadratic-time methods

- Two-sample, independence, interaction, goodness-of-fit test.
- Kernel endowed domain (goodness-of-fit $\mathbb{R}^d$).
- Typically: null can be ugly, techniques do not scale well.

Next time

Linear-time tests, with high-power!

Questions

- Lancaster-interaction measure: reason of the last term?
- Stein operator: why does it work?
- Stein operator: how to estimate it?

Lancaster interaction

Interaction measure:

$$L(\mathbb{P}) = \mathbb{P} - \mathbb{P}_{1,2} \otimes \mathbb{P}_3 - \mathbb{P}_{2,3} \otimes \mathbb{P}_1 - \mathbb{P}_{1,3} \otimes \mathbb{P}_2 + 2\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3.$$

Assume for example:

$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_{2,3} \quad \Rightarrow \quad \mathbb{P}_{1,2} = \mathbb{P}_1 \otimes \mathbb{P}_2, \quad \mathbb{P}_{1,3} = \mathbb{P}_1 \otimes \mathbb{P}_3,$$

$$x_1 \perp [x_2; x_3], \qquad x_1 \perp x_2, \qquad x_1 \perp x_3,$$

Lancaster interaction

Interaction measure:

$$L(\mathbb{P}) = \mathbb{P} - \mathbb{P}_{1,2} \otimes \mathbb{P}_3 - \mathbb{P}_{2,3} \otimes \mathbb{P}_1 - \mathbb{P}_{1,3} \otimes \mathbb{P}_2 + 2\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3.$$

Assume for example:

$$\begin{array}{ccccc} \mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_{2,3} & \Rightarrow & \mathbb{P}_{1,2} = \mathbb{P}_1 \otimes \mathbb{P}_2, & & \mathbb{P}_{1,3} = \mathbb{P}_1 \otimes \mathbb{P}_3, \\ x_1 \perp [x_2; x_3], & & x_1 \perp x_2, & & x_1 \perp x_3, \end{array}$$

and L simplifies to

$$L(\mathbb{P}) = \mathbb{P} - \underbrace{\mathbb{P}_{1,2} \otimes \mathbb{P}_3}_{\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3} - \mathbb{P}_{2,3} \otimes \mathbb{P}_1 - \underbrace{\mathbb{P}_{1,3} \otimes \mathbb{P}_2}_{\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3} + 2\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_3 = 0.$$

Stein operator ($d = 1$ for simplicity): why?

Let $f \in \mathcal{H}_k$.

$$(S_p f)(x) = [\log p(x)]' f(x) + f'(x)$$

Stein operator ($d = 1$ for simplicity): why?

Let $f \in \mathcal{H}_k$.

$$
\begin{aligned}
(S_p f)(x) &= [\log p(x)]' f(x) + f'(x) \\
&= \frac{p'(x)}{p(x)} f(x) + \frac{f'(x)}{p(x)} p(x)
\end{aligned}
$$

Stein operator ($d = 1$ for simplicity): why?

Let $f \in \mathcal{H}_k$.

$$\begin{aligned}(S_p f)(x) &= [\log p(x)]' f(x) + f'(x) \\ &= \frac{p'(x)}{p(x)} f(x) + \frac{f'(x)}{p(x)} p(x) = \frac{[p(x)f(x)]'}{p(x)}.\end{aligned}$$

Stein operator ($d = 1$ for simplicity): why?

Let $f \in \mathcal{H}_k$.

$$(S_p f)(x) = [\log p(x)]' f(x) + f'(x)$$
$$= \frac{p'(x)}{p(x)} f(x) + \frac{f'(x)}{p(x)} p(x) = \frac{[p(x)f(x)]'}{p(x)}.$$

$p = q$ implies: for any f

$$\mathbb{E}_{x\sim q}(S_p f)(x) = \int_{\mathbb{R}} \frac{[p(x)f(x)]'}{p(x)} p(x)\,\mathrm{d}x$$

Stein operator ($d = 1$ for simplicity): why?

Let $f \in \mathcal{H}_k$.

$$(S_p f)(x) = [\log p(x)]' f(x) + f'(x)$$
$$= \frac{p'(x)}{p(x)} f(x) + \frac{f'(x)}{p(x)} p(x) = \frac{[p(x)f(x)]'}{p(x)}.$$

$p = q$ implies: for any f

$$\mathbb{E}_{x \sim q}(S_p f)(x) = \int_{\mathbb{R}} \frac{[p(x)f(x)]'}{p(x)} p(x)\,\mathrm{d}x$$
$$= \int_{\mathbb{R}} [p(x)f(x)]'\,\mathrm{d}x$$

Stein operator ($d = 1$ for simplicity): why?

Let $f \in \mathcal{H}_k$.

$$(S_p f)(x) = [\log p(x)]' f(x) + f'(x)$$
$$= \frac{p'(x)}{p(x)} f(x) + \frac{f'(x)}{p(x)} p(x) = \frac{[p(x)f(x)]'}{p(x)}.$$

$p = q$ implies: for any f

$$\mathbb{E}_{x\sim q}(S_p f)(x) = \int_{\mathbb{R}} \frac{[p(x)f(x)]'}{p(x)} p(x)\,\mathrm{d}x$$
$$= \int_{\mathbb{R}} [p(x)f(x)]'\,\mathrm{d}x = [p(x)f(x)]_{x=-\infty}^{x=\infty} = 0.$$

Assumption: $\lim_{|x|\to\infty} p(x)f(x) = 0$.

Stein operator: computation

Test statistics:

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x \sim q}(S_p f)(x).$$

We rewrite $(S_p f)(x)$ by the reproducing property:

$$(S_p f)(x) = [\log p(x)]' f(x) + f'(x)$$

Stein operator: computation

Test statistics:

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x \sim q}(S_p f)(x).$$

We rewrite $(S_p f)(x)$ by the reproducing property:

$$\begin{aligned}(S_p f)(x) &= [\log p(x)]' f(x) + f'(x) \\ &= [\log p(x)]' \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} + \left\langle f, k'(\cdot, x) \right\rangle_{\mathcal{H}_k}\end{aligned}$$

Stein operator: computation

Test statistics:

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x \sim q}(S_p f)(x).$$

We rewrite $(S_p f)(x)$ by the reproducing property:

$$\begin{aligned}
(S_p f)(x) &= [\log p(x)]' f(x) + f'(x) \\
&= [\log p(x)]' \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} + \left\langle f, k'(\cdot, x) \right\rangle_{\mathcal{H}_k} \\
&= \Big\langle f, \underbrace{[\log p(x)]' k(\cdot, x) + k'(\cdot, x)}_{=:\xi_p(\cdot, x)} \Big\rangle_{\mathcal{H}_k}.
\end{aligned}$$

Stein operator: computation

Test statistics:

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x\sim q}(S_p f)(x).$$

We rewrite $(S_p f)(x)$ by the reproducing property:

$$\begin{aligned}
(S_p f)(x) &= [\log p(x)]' f(x) + f'(x) \\
&= [\log p(x)]' \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} + \langle f, k'(\cdot, x)\rangle_{\mathcal{H}_k} \\
&= \Big\langle f, \underbrace{[\log p(x)]' k(\cdot, x) + k'(\cdot, x)}_{=:\xi_p(\cdot, x)} \Big\rangle_{\mathcal{H}_k}.
\end{aligned}$$

Thus,

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x\sim q}\big\langle f, \xi_p(\cdot, x)\big\rangle_{\mathcal{H}_k}$$

Stein operator: computation

Test statistics:

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x \sim q}(S_p f)(x).$$

We rewrite $(S_p f)(x)$ by the reproducing property:

$$\begin{aligned}(S_p f)(x) &= [\log p(x)]' f(x) + f'(x) \\ &= [\log p(x)]' \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} + \langle f, k'(\cdot, x) \rangle_{\mathcal{H}_k} \\ &= \Big\langle f, \underbrace{[\log p(x)]' k(\cdot, x) + k'(\cdot, x)}_{=:\xi_p(\cdot, x)} \Big\rangle_{\mathcal{H}_k}.\end{aligned}$$

Thus,

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x \sim q} \langle f, \xi_p(\cdot, x) \rangle_{\mathcal{H}_k} = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \langle f, \mathbb{E}_{x \sim q} \xi_p(\cdot, x) \rangle_{\mathcal{H}_k}$$

Stein operator: computation

Test statistics:

$$T_p(q) = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x\sim q}(S_p f)(x).$$

We rewrite $(S_p f)(x)$ by the reproducing property:

$$\begin{aligned}(S_p f)(x) &= [\log p(x)]' f(x) + f'(x)\\ &= [\log p(x)]' \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} + \langle f, k'(\cdot, x)\rangle_{\mathcal{H}_k}\\ &= \Big\langle f, \underbrace{[\log p(x)]' k(\cdot, x) + k'(\cdot, x)}_{=:\xi_p(\cdot, x)} \Big\rangle_{\mathcal{H}_k}.\end{aligned}$$

Thus,

$$\begin{aligned}T_p(q) &= \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \mathbb{E}_{x\sim q}\langle f, \xi_p(\cdot, x)\rangle_{\mathcal{H}_k} = \sup_{\|f\|_{\mathcal{H}_k} \leqslant 1} \langle f, \mathbb{E}_{x\sim q}\xi_p(\cdot, x)\rangle_{\mathcal{H}_k}\\ &= \|\mathbb{E}_{x\sim q}\xi_p(\cdot, x)\|_{\mathcal{H}_k} =: \|g\|_{\mathcal{H}_k}, \; g: \text{Stein witness function.}\end{aligned}$$

Stein operator: computation finished

Until now: with $g = \mathbb{E}_{x\sim q}\xi_p(\cdot, x)$, $\xi_p(\cdot, x) = [\log p(x)]' k(\cdot, x) + k'(\cdot, x)$

$$[T_p(q)]^2 = \|g\|^2_{\mathcal{H}_k} = \left\langle \mathbb{E}_{x\sim q}\xi_p(\cdot, x), \mathbb{E}_{x'\sim q}\xi_p(\cdot, x') \right\rangle_{\mathcal{H}_k}$$

Stein operator: computation finished

Until now: with $g = \mathbb{E}_{x\sim q}\xi_p(\cdot,x)$, $\xi_p(\cdot,x) = [\log p(x)]' k(\cdot,x) + k'(\cdot,x)$

$$[T_p(q)]^2 = \|g\|^2_{\mathcal{H}_k} = \left\langle \mathbb{E}_{x\sim q}\xi_p(\cdot,x), \mathbb{E}_{x'\sim q}\xi_p(\cdot,x') \right\rangle_{\mathcal{H}_k}$$
$$= \mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}\left\langle \xi_p(\cdot,x), \xi_p(\cdot,x') \right\rangle_{\mathcal{H}_k},$$

Stein operator: computation finished

Until now: with $g = \mathbb{E}_{x\sim q}\xi_p(\cdot, x)$, $\xi_p(\cdot, x) = [\log p(x)]' k(\cdot, x) + k'(\cdot, x)$

$$[T_p(q)]^2 = \|g\|^2_{\mathcal{H}_k} = \left\langle \mathbb{E}_{x\sim q}\xi_p(\cdot, x), \mathbb{E}_{x'\sim q}\xi_p(\cdot, x') \right\rangle_{\mathcal{H}_k}$$
$$= \mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}\left\langle \xi_p(\cdot, x), \xi_p(\cdot, x') \right\rangle_{\mathcal{H}_k},$$
$$\left\langle \xi_p(\cdot, x), \xi_p(\cdot, x') \right\rangle_{\mathcal{H}_k} = \left\langle [\log p(x)]' k(\cdot, x) + k'(\cdot, x), [\log p(x')]' k(\cdot, x') + k'(\cdot, x') \right\rangle_{\mathcal{H}_k} =: h_p(x, x'),$$

Stein operator: computation finished

Until now: with $g = \mathbb{E}_{x\sim q}\xi_p(\cdot,x)$, $\xi_p(\cdot,x) = [\log p(x)]'k(\cdot,x) + k'(\cdot,x)$

$$
\begin{aligned}
[T_p(q)]^2 &= \|g\|^2_{\mathcal{H}_k} = \left\langle \mathbb{E}_{x\sim q}\xi_p(\cdot,x), \mathbb{E}_{x'\sim q}\xi_p(\cdot,x')\right\rangle_{\mathcal{H}_k} \\
&= \mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}\left\langle \xi_p(\cdot,x), \xi_p(\cdot,x')\right\rangle_{\mathcal{H}_k}, \\
\left\langle \xi_p(\cdot,x), \xi_p(\cdot,x')\right\rangle_{\mathcal{H}_k} &= \left\langle [\log p(x)]'k(\cdot,x) + k'(\cdot,x),\right. \\
&\qquad \left. [\log p(x')]'k(\cdot,x') + k'(\cdot,x')\right\rangle_{\mathcal{H}_k} =: h_p(x,x'), \\
h_p(x,y) &= [\log p(x)]'[\log p(y)]'k(x,y) + [\log p(x)]'k'_y(x,y) + \\
&\quad [\log p(y)]'k'_x(x,y) + k''_{xy}(x,y).
\end{aligned}
$$

Stein operator: computation finished

Until now: with $g = \mathbb{E}_{x\sim q}\xi_p(\cdot,x)$, $\xi_p(\cdot,x) = [\log p(x)]'k(\cdot,x) + k'(\cdot,x)$

$$
\begin{aligned}
[T_p(q)]^2 &= \|g\|^2_{\mathcal{H}_k} = \left\langle \mathbb{E}_{x\sim q}\xi_p(\cdot,x), \mathbb{E}_{x'\sim q}\xi_p(\cdot,x') \right\rangle_{\mathcal{H}_k} \\
&= \mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}\left\langle \xi_p(\cdot,x), \xi_p(\cdot,x') \right\rangle_{\mathcal{H}_k}, \\
\left\langle \xi_p(\cdot,x), \xi_p(\cdot,x') \right\rangle_{\mathcal{H}_k} &= \left\langle [\log p(x)]'k(\cdot,x) + k'(\cdot,x), \right. \\
&\qquad \left. [\log p(x')]'k(\cdot,x') + k'(\cdot,x') \right\rangle_{\mathcal{H}_k} =: h_p(x,x'), \\
h_p(x,y) &= [\log p(x)]'[\log p(y)]'k(x,y) + [\log p(x)]'k'_y(x,y) + \\
&\qquad [\log p(y)]'k'_x(x,y) + k''_{xy}(x,y).
\end{aligned}
$$

$\Rightarrow$ Quadratic-time estimator (U-statistic):

$$
\widehat{[T_p(q)]^2} = \frac{1}{n(n-1)}\sum_{i\neq j} h_p(x_i,x_j).
$$

Hypothesis testing: linear-time methods

Outline

- Nyström method, random Fourier features.
- Analytic representations → linear-time two-sample testing.
- High-power linear-time techniques:
  - two-sample testing,
  - independence testing.

Three schemes

Exemplified in independence testing [Zhang et al., 2017]:

- block-HSIC: analog of block-MMD.

Three schemes

Exemplified in independence testing [Zhang et al., 2017]:

- block-HSIC: analog of block-MMD.
- 2 low-rank schemes:

Three schemes

Exemplified in independence testing [Zhang et al., 2017]:

- block-HSIC: analog of block-MMD.
- 2 low-rank schemes:
  - Nyström method
    [Williams and Seeger, 2001, Drineas and Mahoney, 2005].

Three schemes

Exemplified in independence testing [Zhang et al., 2017]:

- block-HSIC: analog of block-MMD.
- 2 low-rank schemes:
  - Nyström method [Williams and Seeger, 2001, Drineas and Mahoney, 2005].
  - random Fourier features: [Rahimi and Recht, 2007, Sutherland and Schneider, 2015, Sriperumbudur and Szabó, 2015].

HSIC recall

$$C^c_{xy} = \mathbb{E}_{xy}\Big[\big(\varphi(x) - \mu_x\big) \otimes \big(\psi(y) - \mu_y\big)\Big]$$
$$= \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right] - \mu_x \otimes \mu_y,$$
$$HSIC(x, y) = \left\| C^c_{xy} \right\|_{HS}.$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n \times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n}$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n \times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n \times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$
$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T}$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n\times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$
$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T} = \Phi^{u}\left(\Phi^{u}\right)^{T}, \quad \Phi^{u} \in \mathbb{R}^{n\times r},$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n \times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$

$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T} = \Phi^{u}\left(\Phi^{u}\right)^{T}, \quad \Phi^{u} \in \mathbb{R}^{n \times r},$$

$$\mathbb{R}^{r \times r} \ni \mathbf{C}^{u} = \left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T}\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n \times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$

$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T} = \Phi^{u}\left(\Phi^{u}\right)^{T}, \quad \Phi^{u} \in \mathbb{R}^{n \times r},$$

$$\mathbb{R}^{r \times r} \ni \mathbf{C}^{u} = \left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T}\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}} =: \left(\Phi^{u}\right)^{T}\Phi^{u}.$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n \times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$

$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T} = \Phi^{u}\left(\Phi^{u}\right)^{T}, \quad \Phi^{u} \in \mathbb{R}^{n \times r},$$

$$\mathbb{R}^{r \times r} \ni \mathbf{C}^{u} = \left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T}\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}} =: \left(\Phi^{u}\right)^{T}\Phi^{u}.$$

- With centering:

$$\mathbb{R}^{n \times n} \ni \tilde{\hat{\mathbf{G}}} = \mathbf{H}_n\hat{\mathbf{G}}\mathbf{H}_n$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n\times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n\times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$

$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T} = \Phi^{u}\left(\Phi^{u}\right)^{T}, \quad \Phi^{u} \in \mathbb{R}^{n\times r},$$

$$\mathbb{R}^{r\times r} \ni \mathbf{C}^{u} = \left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T}\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}} =: \left(\Phi^{u}\right)^{T}\Phi^{u}.$$

- With centering:

$$\mathbb{R}^{n\times n} \ni \tilde{\hat{\mathbf{G}}} = \mathbf{H}_{n}\hat{\mathbf{G}}\mathbf{H}_{n} = \mathbf{H}_{n}\Phi^{u}\left(\Phi^{u}\right)^{T}\mathbf{H}_{n},$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n \times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n\times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^T$$

$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^T = \Phi^u\left(\Phi^u\right)^T, \quad \Phi^u \in \mathbb{R}^{n\times r},$$

$$\mathbb{R}^{r\times r} \ni \mathbf{C}^u = \left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^T \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}} =: \left(\Phi^u\right)^T\Phi^u.$$

- With centering:

$$\mathbb{R}^{n\times n} \ni \tilde{\hat{\mathbf{G}}} = \mathbf{H}_n\hat{\mathbf{G}}\mathbf{H}_n = \mathbf{H}_n\Phi^u\left(\Phi^u\right)^T\mathbf{H}_n,$$

$$\mathbb{R}^{r\times r} \ni \mathbf{C}^c = \left(\Phi^u\right)^T\mathbf{H}_n\mathbf{H}_n\Phi^u$$

Nyström method

Idea

Approximate $\mathbf{G} \in \mathbb{R}^{n\times n}$ with a (random) subset of size $r \ll n$.

- Without centering:

$$\mathbb{R}^{n\times n} \ni \hat{\mathbf{G}} \approx \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-1}\mathbf{G}_{r,n} = \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{r,r}^{-\frac{1}{2}}\mathbf{G}_{n,r}^{T}$$

$$= \mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T} = \Phi^{u}\left(\Phi^{u}\right)^{T}, \quad \Phi^{u} \in \mathbb{R}^{n\times r},$$

$$\mathbb{R}^{r\times r} \ni \mathbf{C}^{u} = \left[\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}}\right]^{T}\mathbf{G}_{n,r}\mathbf{G}_{r,r}^{-\frac{1}{2}} =: \left(\Phi^{u}\right)^{T}\Phi^{u}.$$

- With centering:

$$\mathbb{R}^{n\times n} \ni \tilde{\hat{\mathbf{G}}} = \mathbf{H}_{n}\hat{\mathbf{G}}\mathbf{H}_{n} = \mathbf{H}_{n}\Phi^{u}\left(\Phi^{u}\right)^{T}\mathbf{H}_{n},$$

$$\mathbb{R}^{r\times r} \ni \mathbf{C}^{c} = \left(\Phi^{u}\right)^{T}\mathbf{H}_{n}\mathbf{H}_{n}\Phi^{u} =: \left(\Phi^{c}\right)^{T}\Phi^{c}.$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T \Phi_x^u$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = \left(\Phi_x^u\right)^T \Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T; \ldots; \left(\Phi_{x,n}^u\right)^T \right] \in \mathbb{R}^{n \times r_x},$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T \Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T; \ldots; \left(\Phi_{x,n}^u\right)^T \right] \in \mathbb{R}^{n \times r_x},$$

$$\tilde{\hat{\mathbf{G}}}_x = \mathbf{H}_n \hat{\mathbf{G}}_x \mathbf{H}_n$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T \Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T; \ldots; \left(\Phi_{x,n}^u\right)^T \right] \in \mathbb{R}^{n \times r_x},$$

$$\tilde{\hat{\mathbf{G}}}_x = \mathbf{H}_n \hat{\mathbf{G}}_x \mathbf{H}_n = \mathbf{H}_n \Phi_x^u \left(\Phi_x^u\right)^T \mathbf{H}_n$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T \Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T ; \ldots ; \left(\Phi_{x,n}^u\right)^T \right] \in \mathbb{R}^{n \times r_x},$$

$$\tilde{\hat{\mathbf{G}}}_x = \mathbf{H}_n \hat{\mathbf{G}}_x \mathbf{H}_n = \mathbf{H}_n \Phi_x^u (\Phi_x^u)^T \mathbf{H}_n, \quad \mathbf{C}_x^c = (\Phi_x^u)^T \mathbf{H}_n \mathbf{H}_n \Phi_x^u$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u(\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T\Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T; \ldots; \left(\Phi_{x,n}^u\right)^T\right] \in \mathbb{R}^{n\times r_x},$$

$$\tilde{\hat{\mathbf{G}}}_x = \mathbf{H}_n\hat{\mathbf{G}}_x\mathbf{H}_n = \mathbf{H}_n\Phi_x^u\left(\Phi_x^u\right)^T\mathbf{H}_n, \quad \mathbf{C}_x^c = \left(\Phi_x^u\right)^T\mathbf{H}_n\mathbf{H}_n\Phi_x^u =: \left(\Phi_x^c\right)^T\Phi_x^c.$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u(\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T\Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T; \ldots; \left(\Phi_{x,n}^u\right)^T\right] \in \mathbb{R}^{n \times r_x},$$

$$\tilde{\hat{\mathbf{G}}}_x = \mathbf{H}_n\hat{\mathbf{G}}_x\mathbf{H}_n = \mathbf{H}_n\Phi_x^u\left(\Phi_x^u\right)^T\mathbf{H}_n, \quad \mathbf{C}_x^c = \left(\Phi_x^u\right)^T\mathbf{H}_n\mathbf{H}_n\Phi_x^u =: \left(\Phi_x^c\right)^T\Phi_x^c.$$

On y:

$$\hat{\mathbf{G}}_y \approx \Phi_y^u(\Phi_y^u)^T \Rightarrow \mathbf{C}_y^u = \left(\Phi_y^u\right)^T\Phi_y^u, \quad \Phi_y^u = \left[\left(\Phi_{y,1}^u\right)^T; \ldots; \left(\Phi_{y,n}^u\right)^T\right] \in \mathbb{R}^{n \times r_y},$$

Implementation for x and y, separately

On x:

$$\hat{\mathbf{G}}_x \approx \Phi_x^u (\Phi_x^u)^T \Rightarrow \mathbf{C}_x^u = (\Phi_x^u)^T \Phi_x^u, \quad \Phi_x^u = \left[\left(\Phi_{x,1}^u\right)^T; \ldots; \left(\Phi_{x,n}^u\right)^T \right] \in \mathbb{R}^{n \times r_x},$$

$$\tilde{\hat{\mathbf{G}}}_x = \mathbf{H}_n \hat{\mathbf{G}}_x \mathbf{H}_n = \mathbf{H}_n \Phi_x^u \left(\Phi_x^u\right)^T \mathbf{H}_n, \quad \mathbf{C}_x^c = \left(\Phi_x^u\right)^T \mathbf{H}_n \mathbf{H}_n \Phi_x^u =: \left(\Phi_x^c\right)^T \Phi_x^c.$$

On y:

$$\hat{\mathbf{G}}_y \approx \Phi_y^u (\Phi_y^u)^T \Rightarrow \mathbf{C}_y^u = \left(\Phi_y^u\right)^T \Phi_y^u, \quad \Phi_y^u = \left[\left(\Phi_{y,1}^u\right)^T; \ldots; \left(\Phi_{y,n}^u\right)^T \right] \in \mathbb{R}^{n \times r_y},$$

$$\tilde{\hat{\mathbf{G}}}_y \approx \mathbf{H}_n \mathbf{G}_y \mathbf{H}_n = \mathbf{H}_n \Phi_y^u \left(\Phi_y^u\right)^T \mathbf{H}_n, \quad \mathbf{C}_y^c = \left(\Phi_y^u\right)^T \mathbf{H}_n \mathbf{H}_n \Phi_y^u =: \left(\Phi_y^c\right)^T \Phi_y^c.$$

Nyström-based HSIC estimator

Population quantity:

$$HSIC^2(x,y) = \left\|\mathbb{E}_{xy}\left[\varphi(x)\otimes\psi(y)\right] - \mu_x\otimes\mu_y\right\|_{HS}^2$$
$$= \left\|\mathbb{E}_{xy}\left[\left(\varphi(x)-\mu_x\right)\otimes\left(\psi(y)-\mu_y\right)\right]\right\|_{HS}^2.$$

Estimator:

$$\widehat{HSIC^2_{b,N}}(x,y) = \left\|\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\left(\Phi^u_{y,i}\right)^T - \left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\right)\left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{y,i}\right)^T\right\|_F^2$$

Nyström-based HSIC estimator

Population quantity:

$$HSIC^2(x,y) = \left\|\mathbb{E}_{xy}\left[\varphi(x)\otimes\psi(y)\right] - \mu_x\otimes\mu_y\right\|_{HS}^2$$
$$= \left\|\mathbb{E}_{xy}\left[\left(\varphi(x)-\mu_x\right)\otimes\left(\psi(y)-\mu_y\right)\right]\right\|_{HS}^2.$$

Estimator:

$$\widehat{HSIC^2_{b,N}}(x,y) = \left\|\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\left(\Phi^u_{y,i}\right)^T - \left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\right)\left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{y,i}\right)^T\right\|_F^2$$
$$= \left\|\frac{1}{n}\left(\Phi^u_x\right)^T\Phi^u_y - \frac{1}{n^2}\left(\Phi^u_x\right)^T\mathbf{1}_n\mathbf{1}_n^T\Phi^u_y\right\|_F^2$$

Nyström-based HSIC estimator

Population quantity:

$$HSIC^2(x,y) = \left\| \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right] - \mu_x \otimes \mu_y \right\|_{HS}^2$$
$$= \left\| \mathbb{E}_{xy}\left[\left(\varphi(x) - \mu_x\right) \otimes \left(\psi(y) - \mu_y\right) \right] \right\|_{HS}^2 .$$

Estimator:

$$\widehat{HSIC^2_{b,N}}(x,y) = \left\| \frac{1}{n} \sum_{i=1}^n \Phi^u_{x,i} \left(\Phi^u_{y,i}\right)^T - \left(\frac{1}{n} \sum_{i=1}^n \Phi^u_{x,i} \right) \left(\frac{1}{n} \sum_{i=1}^n \Phi^u_{y,i} \right)^T \right\|_F^2$$
$$= \left\| \frac{1}{n} \left(\Phi^u_x\right)^T \Phi^u_y - \frac{1}{n^2} \left(\Phi^u_x\right)^T \mathbf{1}_n \mathbf{1}_n^T \Phi^u_y \right\|_F^2$$
$$= \left\| \frac{1}{n} \left(\Phi^u_x\right)^T \underbrace{\left(\mathbf{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n} \right)}_{\mathbf{H}_n = \mathbf{H}_n^T \mathbf{H}_n} \Phi^u_y \right\|_F^2$$

Nyström-based HSIC estimator

Population quantity:

$$HSIC^2(x,y) = \left\|\mathbb{E}_{xy}\left[\varphi(x)\otimes\psi(y)\right] - \mu_x\otimes\mu_y\right\|_{HS}^2$$
$$= \left\|\mathbb{E}_{xy}\left[\left(\varphi(x)-\mu_x\right)\otimes\left(\psi(y)-\mu_y\right)\right]\right\|_{HS}^2.$$

Estimator:

$$\widehat{HSIC^2_{b,N}}(x,y) = \left\|\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\left(\Phi^u_{y,i}\right)^T - \left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\right)\left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{y,i}\right)^T\right\|_F^2$$
$$= \left\|\frac{1}{n}\left(\Phi^u_x\right)^T\Phi^u_y - \frac{1}{n^2}\left(\Phi^u_x\right)^T\mathbf{1}_n\mathbf{1}_n^T\Phi^u_y\right\|_F^2$$
$$= \left\|\frac{1}{n}\left(\Phi^u_x\right)^T\underbrace{\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\right)}_{\mathbf{H}_n=\mathbf{H}_n^T\mathbf{H}_n}\Phi^u_y\right\|_F^2 = \left\|\frac{1}{n}\left(\Phi^c_x\right)^T\Phi^c_y\right\|_F^2.$$

Nyström-based HSIC estimator – conclusion

$$HSIC^2(x,y) = \left\| C_{xy}^c \right\|_{HS}^2,$$

$$\widehat{HSIC^2_{b,N}}(x,y) = \left\| \frac{1}{n} \left(\Phi_x^c\right)^T \Phi_y^c \right\|_F^2.$$

Nyström-based HSIC estimator – conclusion

$$HSIC^2(x,y) = \left\| C_{xy}^c \right\|_{HS}^2,$$

$$\widehat{HSIC^2_{b,N}}(x,y) = \left\| \frac{1}{n} \left(\Phi_x^c\right)^T \Phi_y^c \right\|_F^2.$$

In short

C_{xy}^c changed to $\frac{1}{n}\left(\Phi_x^c\right)^T \Phi_y^c$, with Frobenius norm.

Nyström technique: notes

- Use $\widehat{HSIC_{b,N}}$ in
  - permutation test, or spectral approach.

Nyström technique: notes

- Use $\widehat{HSIC_{b,N}}$ in
  - permutation test, or spectral approach.
- Computational complexity (null approximation):

$$\mathcal{O}\left(n^3\right) \to \mathcal{O}\left(r_x^3 + r_y^3 + \left(r_x^2 + r_y^2\right) n + r_x r_y n\right).$$

In practice: $r_x, r_y \ll n$.

Nyström technique: notes

- Use $\widehat{HSIC_{b,N}}$ in
  - permutation test, or spectral approach.
- Computational complexity (null approximation):

$$\mathcal{O}\left(n^3\right) \to \mathcal{O}\left(r_x^3 + r_y^3 + \left(r_x^2 + r_y^2\right) n + r_x r_y n\right).$$

  In practice: $r_x, r_y \ll n$.
- GP [Snelson and Ghahramani, 2006, Titsias, 2009]:
  - subset $\to$ optimized subset of size r,
  - inducing points.

Random Fourier features

Characteristic functions: quick summary [Sasvári, 2013]

$\mathbb{P} \mapsto \phi_{\mathbb{P}}$:

$$\phi_{\mathbb{P}}(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[e^{i\langle\mathbf{t},\mathbf{x}\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle\mathbf{t},\mathbf{x}\rangle}\mathrm{d}\mathbb{P}(\mathbf{x}), \quad \mathbf{t}\in\mathbb{R}^d.$$

Characteristic functions: quick summary [Sasvári, 2013]

$\mathbb{P} \mapsto \phi_{\mathbb{P}}$:

$$\phi_{\mathbb{P}}(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[e^{i\langle \mathbf{t},\mathbf{x}\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle \mathbf{t},\mathbf{x}\rangle} \mathrm{d}\mathbb{P}(\mathbf{x}), \quad \mathbf{t} \in \mathbb{R}^d.$$

Properties:

- $\exists$, $\mathbb{P} \overset{1:1}{\longleftrightarrow} \phi_{\mathbb{P}}$,

Characteristic functions: quick summary [Sasvári, 2013]

$\mathbb{P} \mapsto \phi_{\mathbb{P}}$:

$$\phi_{\mathbb{P}}(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[e^{i\langle\mathbf{t},\mathbf{x}\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle\mathbf{t},\mathbf{x}\rangle} \mathrm{d}\mathbb{P}(\mathbf{x}), \quad \mathbf{t} \in \mathbb{R}^d.$$

Properties:

- $\exists$, $\mathbb{P} \overset{1:1}{\longleftrightarrow} \phi_{\mathbb{P}}$,
- $|\phi_{\mathbb{P}}(\mathbf{t})| \leqslant 1$, $\phi_{\mathbb{P}}(-\mathbf{t}) = \overline{\phi_{\mathbb{P}}(\mathbf{t})} \quad \forall \mathbf{t} \in \mathbb{R}^d$.

Characteristic functions: quick summary [Sasvári, 2013]

$\mathbb{P} \mapsto \phi_{\mathbb{P}}$:

$$\phi_{\mathbb{P}}(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[e^{i\langle\mathbf{t},\mathbf{x}\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle\mathbf{t},\mathbf{x}\rangle}\mathrm{d}\mathbb{P}(\mathbf{x}), \quad \mathbf{t}\in\mathbb{R}^d.$$

Properties:

- $\exists$, $\mathbb{P} \overset{1:1}{\longleftrightarrow} \phi_{\mathbb{P}}$,
- $|\phi_{\mathbb{P}}(\mathbf{t})| \leqslant 1$, $\phi_{\mathbb{P}}(-\mathbf{t}) = \overline{\phi_{\mathbb{P}}(\mathbf{t})} \quad \forall\mathbf{t}\in\mathbb{R}^d$.
- $\phi_{\mathbb{P}}$: uniformly continuous on $\mathbb{R}^d$.

Characteristic functions: quick summary [Sasvári, 2013]

$\mathbb{P} \mapsto \phi_{\mathbb{P}}$:

$$\phi_{\mathbb{P}}(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[e^{i\langle\mathbf{t},\mathbf{x}\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle\mathbf{t},\mathbf{x}\rangle}\mathrm{d}\mathbb{P}(\mathbf{x}), \quad \mathbf{t} \in \mathbb{R}^d.$$

Properties:

- $\exists$, $\mathbb{P} \overset{1:1}{\longleftrightarrow} \phi_{\mathbb{P}}$,
- $|\phi_{\mathbb{P}}(\mathbf{t})| \leqslant 1$, $\phi_{\mathbb{P}}(-\mathbf{t}) = \overline{\phi_{\mathbb{P}}(\mathbf{t})} \quad \forall \mathbf{t} \in \mathbb{R}^d$.
- $\phi_{\mathbb{P}}$: uniformly continuous on $\mathbb{R}^d$.
- pd: $\sum_{i,j=1}^n \phi_{\mathbb{P}}(\mathbf{t}_i - \mathbf{t}_j)c_i\bar{c}_j \geqslant 0$, for $\forall n \in \mathbb{Z}^+$, $\mathbf{t}_i \in \mathbb{R}^d$, $c_i \in \mathbb{C}$.

Characteristic functions: quick summary [Sasvári, 2013]

$\mathbb{P} \mapsto \phi_{\mathbb{P}}$:

$$\phi_{\mathbb{P}}(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[e^{i\langle \mathbf{t},\mathbf{x}\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle \mathbf{t},\mathbf{x}\rangle} \mathrm{d}\mathbb{P}(\mathbf{x}), \quad \mathbf{t} \in \mathbb{R}^d.$$

Properties:

- $\exists$, $\mathbb{P} \overset{1:1}{\longleftrightarrow} \phi_{\mathbb{P}}$,
- $|\phi_{\mathbb{P}}(\mathbf{t})| \leqslant 1$, $\phi_{\mathbb{P}}(-\mathbf{t}) = \overline{\phi_{\mathbb{P}}(\mathbf{t})} \quad \forall \mathbf{t} \in \mathbb{R}^d$.
- $\phi_{\mathbb{P}}$: uniformly continuous on $\mathbb{R}^d$.
- pd: $\sum_{i,j=1}^{n} \phi_{\mathbb{P}}(\mathbf{t}_i - \mathbf{t}_j) c_i \bar{c}_j \geqslant 0$, for $\forall n \in \mathbb{Z}^+$, $\mathbf{t}_i \in \mathbb{R}^d$, $c_i \in \mathbb{C}$.

Recall

Bochner's theorem & $\mathbf{G} \succcurlyeq 0$ definition of kernels!

Characteristic functions: continued

Operations, closedness:

- Sum of independent variables:

$$\phi_{\sum_{i=1}^{n} \mathbf{x}_i}(\mathbf{t}) = \prod_{i=1}^{n} \phi_{\mathbf{x}_i}(\mathbf{t}), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

Characteristic functions: continued

Operations, closedness:

- Sum of independent variables:

$$\phi_{\sum_{i=1}^{n}\mathbf{x}_i}(\mathbf{t}) = \prod_{i=1}^{n}\phi_{\mathbf{x}_i}(\mathbf{t}), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

- Affine transformation ($\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$):

$$\phi_{\mathbf{A}\mathbf{x}+\mathbf{b}}(\mathbf{t}) = e^{i\langle \mathbf{t},\mathbf{b}\rangle}\phi_{\mathbf{x}}\left(\mathbf{A}^T\mathbf{t}\right), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

Characteristic functions: continued

Operations, closedness:

- Sum of independent variables:

$$\phi_{\sum_{i=1}^n \mathbf{x}_i}(\mathbf{t}) = \prod_{i=1}^n \phi_{\mathbf{x}_i}(\mathbf{t}), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

- Affine transformation ($\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$):

$$\phi_{\mathbf{A}\mathbf{x}+\mathbf{b}}(\mathbf{t}) = e^{i\langle \mathbf{t},\mathbf{b}\rangle}\phi_{\mathbf{x}}\left(\mathbf{A}^T\mathbf{t}\right), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

- Concatenation of independent variables: $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_n]$

$$\phi_{\mathbf{x}}(\mathbf{t}) = \prod_{i=1}^n \phi_{\mathbf{x}_i}(\mathbf{t}_i), \quad \mathbf{t} = [\mathbf{t}_1; \ldots; \mathbf{t}_n] \in \mathbb{R}^d.$$

Characteristic functions: continued

Operations, closedness:

- Sum of independent variables:

$$\phi_{\sum_{i=1}^{n}\mathbf{x}_i}(\mathbf{t}) = \prod_{i=1}^{n}\phi_{\mathbf{x}_i}(\mathbf{t}), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

- Affine transformation ($\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$):

$$\phi_{\mathbf{Ax}+\mathbf{b}}(\mathbf{t}) = e^{i\langle \mathbf{t},\mathbf{b}\rangle}\phi_{\mathbf{x}}\left(\mathbf{A}^T\mathbf{t}\right), \quad \forall \mathbf{t} \in \mathbb{R}^d.$$

- Concatenation of independent variables: $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_n]$

$$\phi_{\mathbf{x}}(\mathbf{t}) = \prod_{i=1}^{n}\phi_{\mathbf{x}_i}(\mathbf{t}_i), \quad \mathbf{t} = [\mathbf{t}_1; \ldots; \mathbf{t}_n] \in \mathbb{R}^d.$$

Recall

Distance covariance!

Characteristic functions: continued

Moment condition on $\mathbb{P} \Rightarrow$ differentiability of $\phi_{\mathbb{P}}$.

Assume that exists:

$$M_{\mathbf{a}} = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\left[\mathbf{x}^{\mathbf{a}}\right] \quad \mathbf{a} \in \mathbb{N}^d, \ \left(\mathbf{x}^{\mathbf{a}} := \prod_{i=1}^{d} x_i^{a_i}\right).$$

Then $\exists \partial^{\mathbf{a}} \phi_{\mathbb{P}}$ and

$$\partial^{\mathbf{a}} \phi_{\mathbb{P}}(\mathbf{t}) = i^{|\mathbf{a}|} \int_{\mathbb{R}^d} \mathbf{x}^{\mathbf{a}} e^{i\langle \mathbf{t},\mathbf{x}\rangle} \mathrm{d}\mathbb{P}(x), \ \forall \mathbf{t} \in \mathbb{R}^d,$$
$$\partial^{\mathbf{a}} \phi_{\mathbb{P}}(\mathbf{0}) = i^{|\mathbf{a}|} M_{\mathbf{a}},$$

and $\partial^{\mathbf{a}} \phi_{\mathbb{P}}$ is uniformly continuous.

RFF idea

- k: continuous, shift-invariant on $\mathbb{R}^d$ $[k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y})]$. By Bochner:

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})}}_{\cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))+i\sin(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))} \mathrm{d}\Lambda(\boldsymbol{\omega})$$

RFF idea

- k: continuous, shift-invariant on $\mathbb{R}^d$ $[k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y})]$. By Bochner:

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})}}_{\cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))+i\sin(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))} \mathrm{d}\Lambda(\boldsymbol{\omega})$$
$$= \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

RFF idea

- k: continuous, shift-invariant on $\mathbb{R}^d$ $[k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y})]$. By Bochner:

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})}}_{\cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))+i\sin(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))} \mathrm{d}\Lambda(\boldsymbol{\omega})$$
$$= \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

- RFF trick [Rahimi and Recht, 2007] (MC): $\boldsymbol{\omega}_{1:m} := (\boldsymbol{\omega}_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$,

$$\hat{k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^m \cos\left(\boldsymbol{\omega}_j^T(\mathbf{x}-\mathbf{y})\right)$$

RFF idea

- k: continuous, shift-invariant on $\mathbb{R}^d$ $[k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y})]$. By Bochner:

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})}}_{\cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))+i\sin(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))} \mathrm{d}\Lambda(\boldsymbol{\omega})$$

$$= \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

- RFF trick [Rahimi and Recht, 2007] (MC): $\boldsymbol{\omega}_{1:m} := (\boldsymbol{\omega}_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$,

$$\hat{k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^m \cos\left(\boldsymbol{\omega}_j^T(\mathbf{x}-\mathbf{y})\right) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda_m(\boldsymbol{\omega}).$$

RFF idea

- k: continuous, shift-invariant on $\mathbb{R}^d$ $[k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y})]$. By Bochner:

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})}}_{\cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))+i\sin(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y}))} \mathrm{d}\Lambda(\boldsymbol{\omega})$$

$$= \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

- RFF trick [Rahimi and Recht, 2007] (MC): $\boldsymbol{\omega}_{1:m} := (\boldsymbol{\omega}_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$,

$$\hat{k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^m \cos\left(\boldsymbol{\omega}_j^T(\mathbf{x}-\mathbf{y})\right) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda_m(\boldsymbol{\omega}).$$

Remember (characteristic kernels)

We saw many $k \to \Lambda$ examples!

Questions

- Why is RFF useful?
- Does it converge ($k - \hat{k}$)? Rates?

Why is RFF useful?

Kernel approximation:

$$\hat{k}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{j=1}^{m} \cos\left(\omega_j^T (\mathbf{x} - \mathbf{y})\right).$$

Why is RFF useful?

Kernel approximation:

$$\hat{k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^{m}\cos\left(\omega_j^T(\mathbf{x}-\mathbf{y})\right).$$

By the trigonometric identity:

$$\cos(a-b) = \cos(a)\cos(b) + \sin(a)\sin(b),$$

Why is RFF useful?

Kernel approximation:

$$\hat{k}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{j=1}^{m} \cos\left(\boldsymbol{\omega}_j^T (\mathbf{x} - \mathbf{y})\right).$$

By the trigonometric identity:

$$\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b),$$

$$\hat{k}(\mathbf{x}, \mathbf{y}) = \left\langle \hat{\phi}(\mathbf{x}), \hat{\phi}(\mathbf{y}) \right\rangle_{\mathbb{R}^{2m}},$$

$$\hat{\phi}(\mathbf{x}) = \frac{1}{\sqrt{m}} \left[\cos\left(\boldsymbol{\omega}_1^T \mathbf{x}\right); \ldots; \cos\left(\boldsymbol{\omega}_m^T \mathbf{x}\right); \right.$$
$$\left. \sin\left(\boldsymbol{\omega}_1^T \mathbf{x}\right); \ldots; \sin\left(\boldsymbol{\omega}_m^T \mathbf{x}\right)\right] \in \mathbb{R}^{2m}.$$

Why is RFF useful?

Kernel approximation:

$$\hat{k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^{m} \cos\left(\boldsymbol{\omega}_j^T(\mathbf{x}-\mathbf{y})\right).$$

By the trigonometric identity:

$$\begin{aligned}
\cos(a-b) &= \cos(a)\cos(b) + \sin(a)\sin(b),\\
\hat{k}(\mathbf{x},\mathbf{y}) &= \left\langle \hat{\phi}(\mathbf{x}), \hat{\phi}(\mathbf{y}) \right\rangle_{\mathbb{R}^{2m}},\\
\hat{\phi}(\mathbf{x}) &= \frac{1}{\sqrt{m}}\Big[\cos\left(\boldsymbol{\omega}_1^T\mathbf{x}\right);\ldots;\cos\left(\boldsymbol{\omega}_m^T\mathbf{x}\right);\\
&\qquad \sin\left(\boldsymbol{\omega}_1^T\mathbf{x}\right);\ldots;\sin\left(\boldsymbol{\omega}_m^T\mathbf{x}\right)\Big] \in \mathbb{R}^{2m}.
\end{aligned}$$

Key

We got (random) explicit feature maps!

RFF application in independence testing

Previous slide ⇒

$$\left(\Phi_x^u\right)^T := \left[\hat{\phi}(x_1); \ldots ; \hat{\phi}(x_n)\right], \left(\Phi_y^u\right)^T := \left[\hat{\phi}(y_1); \ldots ; \hat{\phi}(y_n)\right],$$

$$\mathbf{G}_x \approx \Phi_x^u \left(\Phi_x^u\right)^T, \qquad \mathbf{G}_y \approx \Phi_y^u \left(\Phi_y^u\right)^T,$$

RFF application in independence testing

Previous slide $\Rightarrow$

$$\left(\Phi_x^u\right)^T := \left[\hat{\phi}(x_1); \ldots; \hat{\phi}(x_n)\right], \left(\Phi_y^u\right)^T := \left[\hat{\phi}(y_1); \ldots; \hat{\phi}(y_n)\right],$$
$$\mathbf{G}_x \approx \Phi_x^u \left(\Phi_x^u\right)^T, \qquad \mathbf{G}_y \approx \Phi_y^u \left(\Phi_y^u\right)^T,$$

and hence

$$\widehat{HSIC}^2_{b,RFF}(x,y) = \left\| \frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i} \left(\Phi^u_{y,i}\right)^T - \left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{x,i}\right)\left(\frac{1}{n}\sum_{i=1}^n \Phi^u_{y,i}\right)^T \right\|_F^2$$

RFF application in independence testing

Previous slide ⇒

$$\left(\Phi_x^u\right)^T := \left[\hat{\phi}(x_1); \ldots; \hat{\phi}(x_n)\right], \left(\Phi_y^u\right)^T := \left[\hat{\phi}(y_1); \ldots; \hat{\phi}(y_n)\right],$$

$$\mathbf{G}_x \approx \Phi_x^u \left(\Phi_x^u\right)^T, \qquad \mathbf{G}_y \approx \Phi_y^u \left(\Phi_y^u\right)^T,$$

and hence

$$\widehat{HSIC}^2_{b,RFF}(x,y) = \left\| \frac{1}{n} \sum_{i=1}^n \Phi_{x,i}^u \left(\Phi_{y,i}^u\right)^T - \left(\frac{1}{n} \sum_{i=1}^n \Phi_{x,i}^u\right) \left(\frac{1}{n} \sum_{i=1}^n \Phi_{y,i}^u\right)^T \right\|_F^2$$

$$= \ldots = \left\| \frac{1}{n} \left(\Phi_x^c\right)^T \Phi_y^c \right\|_F^2.$$

Briefly

We simply 'overloaded' the features with the RFF ones.

Some further RFF-accelerated measures

- KCCA [Lopez-Paz et al., 2014].
- MMD [Sutherland and Schneider, 2015, Zhao and Meng, 2015, Lopez-Paz, 2016].

RFF: in kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^{\ell}$.
- Task: find $f \in \mathcal{H}_k$ s.t. $f(x_i) \approx y_i$,

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) - y_i]^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \to \min_{f \in \mathcal{H}_k} \quad (\lambda > 0).$$

RFF: in kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^{\ell}$.
- Task: find $f \in \mathcal{H}_k$ s.t. $f(x_i) \approx y_i$,

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) - y_i]^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \to \min_{f \in \mathcal{H}_k} \quad (\lambda > 0).$$

- Analytical solution, $\mathcal{O}(\ell^3)$ – expensive:

$$f(x) = [k(x_1, x), \ldots, k(x_\ell, x)](\mathbf{G} + \lambda \ell I)^{-1}[y_1; \ldots; y_\ell],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{\ell}.$$

RFF: in kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^{\ell}$.
- Task: find $f \in \mathcal{H}_k$ s.t. $f(x_i) \approx y_i$,

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) - y_i]^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \to \min_{f \in \mathcal{H}_k} \quad (\lambda > 0).$$

- Analytical solution, $\mathcal{O}(\ell^3)$ – expensive:

$$f(x) = [k(x_1, x), \ldots, k(x_\ell, x)](\mathbf{G} + \lambda \ell I)^{-1}[y_1; \ldots; y_\ell],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{\ell}.$$

- **Idea**: $\hat{\mathbf{G}}$, matrix-inversion lemma, fast primal solvers → RFF.

Approximation quality

- Hoeffding inequality + union bound [Rahimi and Recht, 2007, Sutherland and Schneider, 2015]:

$$\left\| k - \hat{k} \right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_p\left(|\mathcal{S}| \frac{\sqrt{\log(m)}}{\sqrt{m}} \right).$$

Approximation quality

- Hoeffding inequality + union bound [Rahimi and Recht, 2007, Sutherland and Schneider, 2015]:

$$\left\|k - \hat{k}\right\|_{L^{\infty}(\mathcal{S})} = \mathcal{O}_p\left(|\mathcal{S}|\frac{\sqrt{\log(m)}}{\sqrt{m}}\right).$$

- ECFs [Csörgo and Totik, 1983]: $|\mathcal{S}_m| = e^{o(m)}$ – optimal rate, asymptotic!

Approximation quality

- Hoeffding inequality + union bound [Rahimi and Recht, 2007, Sutherland and Schneider, 2015]:

$$\left\| k - \hat{k} \right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_p\left(|\mathcal{S}| \frac{\sqrt{\log(m)}}{\sqrt{m}} \right).$$

- ECFs [Csörgo and Totik, 1983]: $|\mathcal{S}_m| = e^{o(m)}$ – optimal rate, asymptotic!
- Finite-sample L^∞-bound [Sriperumbudur and Szabó, 2015] $\xrightarrow{\text{spec.}}$

$$\left\| k - \hat{k} \right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_{a.s.}\left(\frac{\sqrt{\log |\mathcal{S}|}}{\sqrt{m}} \right).$$

Approximation quality

- Hoeffding inequality + union bound [Rahimi and Recht, 2007, Sutherland and Schneider, 2015]:

$$\left\| k - \hat{k} \right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_p\left(|\mathcal{S}| \frac{\sqrt{\log(m)}}{\sqrt{m}} \right).$$

- ECFs [Csörgo and Totik, 1983]: $|\mathcal{S}_m| = e^{o(m)}$ – optimal rate, asymptotic!
- Finite-sample L^∞-bound [Sriperumbudur and Szabó, 2015] $\xrightarrow{\text{spec.}}$

$$\left\| k - \hat{k} \right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_{a.s.}\left(\frac{\sqrt{\log |\mathcal{S}|}}{\sqrt{m}} \right).$$

- RFF in ridge regression [Rudi and Rosasco, 2017].

Optimal $\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

- Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$; $g(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right)$]:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| \qquad .$$

Optimal $\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

- Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$; $g(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right)$]:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left|k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y})\right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| =: \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

Optimal $\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

- Empirical process form $[\mathbb{P}g := \int g \mathrm{d}\mathbb{P};\ g(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right)]$:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| =: \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

- $f(\boldsymbol{\omega}_{1:m}) = \|\Lambda - \Lambda_m\|_{\mathcal{G}}$ concentrates (bounded difference):

$$\|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} + \frac{1}{\sqrt{m}}.$$

Optimal $\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

- Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$; $g(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right)$]:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| =: \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

- $f(\boldsymbol{\omega}_{1:m}) = \|\Lambda - \Lambda_m\|_{\mathcal{G}}$ concentrates (bounded difference):

$$\|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} + \frac{1}{\sqrt{m}}.$$

- $\mathcal{G}$ is 'nice' (uniformly bounded, separable Carathéodory) $\Rightarrow$

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \underbrace{\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right)}_{\mathbb{E}_{\epsilon} \sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{j=1}^m \epsilon_j g(\omega_j)\right|}.$$

Proof idea – continued

- Using Dudley's entropy bound:

$$\mathcal{R}\left(\mathcal{G}, \omega_{1:m}\right) \lesssim \frac{1}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)} \mathrm{d}r.$$

Proof idea – continued

- Using Dudley's entropy bound:

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \lesssim \frac{1}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right)} \mathrm{d}r.$$

- $\mathcal{G}$ is smoothly parameterized by a compact set $\Rightarrow$

$$\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right) \leqslant \left(\frac{4|\mathcal{S}|A}{r} + 1\right)^d, \qquad A(\boldsymbol{\omega}_{1:m}) = \sqrt{\frac{1}{m}\sum_{j=1}^m \|\boldsymbol{\omega}_j\|_2^2}.$$

Proof idea – continued

- Using Dudley's entropy bound:

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \lesssim \frac{1}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right)} \mathrm{d}r.$$

- $\mathcal{G}$ is smoothly parameterized by a compact set $\Rightarrow$

$$\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right) \leqslant \left(\frac{4|\mathcal{S}|A}{r} + 1\right)^d, \qquad A(\boldsymbol{\omega}_{1:m}) = \sqrt{\frac{1}{m}\sum_{j=1}^m \|\boldsymbol{\omega}_j\|_2^2}.$$

- Putting together [$|\mathcal{G}|_{L^2(\Lambda_m)} \leqslant 2$, Jensen inequality] we get ...

Theorem (Finite-sample, asymptotically optimal uniform bound on RFF)

Let k be continuous, bounded, shift-invariant, and $\sigma^2 := \int \|\boldsymbol{\omega}\|^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$. Then for $\forall \tau > 0$ and compact set $\mathcal{S} \subset \mathbb{R}^d$

$$\Lambda^m \left(\|\hat{k} - k\|_{L^\infty(\mathcal{S})} \geqslant \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right) \leqslant e^{-\tau},$$

$$h(d, |\mathcal{S}|, \sigma) := 32\sqrt{2d \log(2|\mathcal{S}| + 1)} + 16\sqrt{\frac{2d}{\log(2|\mathcal{S}| + 1)}} + 32\sqrt{2d \log(\sigma + 1)}.$$

Empirical process theory: motivation

The object of interest:

$$\sup_{f \in \mathcal{F}} \left| \mathbb{P}f - \mathbb{P}_n f \right|.$$

Empirical process theory: motivation

The object of interest:

$$\sup_{f \in \mathcal{F}} \left| \mathbb{P} f - \mathbb{P}_n f \right|.$$

Original motivation:

- F: cdf, F_n: empirical cdf.

Empirical process theory: motivation

The object of interest:

$$\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f|.$$

Original motivation:

- F: cdf, F_n: empirical cdf.
- Glivenko-Cantelli theorem:

$$\|F - F_n\|_\infty = \sup_x |F(x) - F_n(x)|$$

Empirical process theory: motivation

The object of interest:

$$\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f|.$$

Original motivation:

- F: cdf, F_n: empirical cdf.
- Glivenko-Cantelli theorem:

$$\|F - F_n\|_\infty = \sup_x |F(x) - F_n(x)|$$
$$= \sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f|, \quad \mathcal{F} = \{\chi_{(\infty,x)} : x \in \mathbb{R}^d\}.$$

Ref: [van der Vaart and Wellner, 1996, van der Vaart, 1998, van de Geer, 2009].

Notes on RFF: L^p bounds, kernel derivatives

- One can also get:
  - $L^p(\mathcal{S})$ results ($\Leftarrow$ uniform bound, type of L^p).
  - bounds on $\partial k^{\mathbf{p},\mathbf{q}}$.

Notes on RFF: L^p bounds, kernel derivatives

- One can also get:
  - $L^p(\mathcal{S})$ results ($\Leftarrow$ uniform bound, type of L^p).
  - bounds on $\partial k^{\mathbf{p},\mathbf{q}}$.
- Kernel derivatives: $\frac{\partial^{\mathbf{p},\mathbf{q}} f(\mathbf{x},\mathbf{y})}{\partial^{\mathbf{p}}\mathbf{x}\partial^{\mathbf{q}}\mathbf{y}}$,
  - nonlinear variable selection [Rosasco et al., 2010, Rosasco et al., 2013],

Notes on RFF: L^p bounds, kernel derivatives

- One can also get:
  - $L^p(\mathcal{S})$ results ($\Leftarrow$ uniform bound, type of L^p).
  - bounds on $\partial k^{\mathbf{p},\mathbf{q}}$.
- Kernel derivatives: $\frac{\partial^{\mathbf{p},\mathbf{q}} f(\mathbf{x},\mathbf{y})}{\partial^{\mathbf{p}}\mathbf{x}\partial^{\mathbf{q}}\mathbf{y}}$,
  - nonlinear variable selection [Rosasco et al., 2010, Rosasco et al., 2013],
  - infinite-dimensional exponential family fitting [Sriperumbudur et al., 2017].

Let us look at the examples!

Nonlinear variable selection

- Objective function, $\lambda > 0$:

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - y_i\right]^2 + \lambda \sum_{j=1}^{d} \|\partial_j f\| \rightarrow \min_{f\in\mathcal{H}_k},$$

$$\|g\| := \sqrt{\frac{1}{n}\sum_{i=1}^{n}|g(x_i)|^2}.$$

Nonlinear variable selection

- Objective function, $\lambda > 0$:

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - y_i\right]^2 + \lambda \sum_{j=1}^{d} \|\partial_j f\| \rightarrow \min_{f \in \mathcal{H}_k},$$

$$\|g\| := \sqrt{\frac{1}{n}\sum_{i=1}^{n} |g(x_i)|^2}.$$

- Intuition:
  - if f does not depend on variable j, then $\partial_j f = 0$.

Infinite-dimensional exponential family ($\mathbb{R}^d$)

- Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, T(\mathbf{x}) \rangle},$$

where $\boldsymbol{\theta}$: natural parameter, $T(\mathbf{x})$: sufficient statistics.

Infinite-dimensional exponential family ($\mathbb{R}^d$)

- Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, T(\mathbf{x}) \rangle},$$

  where $\boldsymbol{\theta}$: natural parameter, $T(\mathbf{x})$: sufficient statistics.
- Examples: normal, exponential, gamma, χ^2, beta, ...

Infinite-dimensional exponential family ($\mathbb{R}^d$)

- Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, T(\mathbf{x}) \rangle},$$

where $\boldsymbol{\theta}$: natural parameter, $T(\mathbf{x})$: sufficient statistics.

- Examples: normal, exponential, gamma, χ^2, beta, ...
- InfiniteD generalization:

$$p_f(\mathbf{x}) \propto e^{f(\mathbf{x})} = e^{\langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k}}.$$

Infinite-dimensional exponential family ($\mathbb{R}^d$)

- Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, T(\mathbf{x}) \rangle},$$

where $\boldsymbol{\theta}$: natural parameter, $T(\mathbf{x})$: sufficient statistics.

- Examples: normal, exponential, gamma, χ^2, beta, ...
- InfiniteD generalization:

$$p_f(\mathbf{x}) \propto e^{f(\mathbf{x})} = e^{\langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k}}.$$

Fitting idea (score matching, Fischer divergence):

$$J(p_*, p_f) := \int p^*(\mathbf{x}) \left\| \frac{\partial \log p_*(\mathbf{x})}{\partial \mathbf{x}} - \frac{\partial \log p_f(\mathbf{x})}{\partial \mathbf{x}} \right\|_2^2 \mathrm{d}\mathbf{x} \to \min_{f \in \mathcal{H}_k}.$$

Notes on RFF: operator-valued extension

- Standard setup: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\mathcal{H}_k = \{f : \mathcal{X} \to \mathbb{R} \mid \ldots\}.$$

Notes on RFF: operator-valued extension

- Standard setup: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\mathcal{H}_k = \{f : \mathcal{X} \to \mathbb{R} \mid \ldots\}.$$

- Operator-valued case:

$$\mathcal{H}_k = \{f : \mathcal{X} \to Y \mid \ldots\}$$

Notes on RFF: operator-valued extension

- Standard setup: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\mathcal{H}_k = \{f : \mathcal{X} \to \mathbb{R} \mid \ldots\}.$$

- Operator-valued case:

$$\mathcal{H}_k = \{f : \mathcal{X} \to Y \mid \ldots\}, \qquad k : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(Y).$$

Y: (separable) Hilbert. Example: $Y = \mathbb{R}^d$, $\mathcal{L}(Y) = \mathbb{R}^{d \times d}$.

Notes on RFF: operator-valued extension

- Standard setup: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\mathcal{H}_k = \{f : \mathcal{X} \to \mathbb{R} \mid \ldots\}.$$

- Operator-valued case:

$$\mathcal{H}_k = \{f : \mathcal{X} \to Y \mid \ldots\}, \qquad k : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(Y).$$

  Y: (separable) Hilbert. Example: $Y = \mathbb{R}^d$, $\mathcal{L}(Y) = \mathbb{R}^{d \times d}$.

- RFF idea
  - works [Brault et al., 2016]; $(\mathbb{R}^d, +) \to$ LCA : ✓
  - open question: 'optimal' rates.

Nyström method, RFF: the end.

Linear-time two-sample testing: analytic representations.

Linear-time 2-sample test [Chwialkowski et al., 2015]

- Recall:

$$MMD(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

Linear-time 2-sample test [Chwialkowski et al., 2015]

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

- Idea: change this to

$$\rho(\mathbb{P},\mathbb{Q}) := \rho\left(\mathbb{P},\mathbb{Q};\{\mathbf{v}_j\}_{j=1}^J\right) := \sqrt{\frac{1}{J}\sum_{j=1}^J [\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}$$

with random $\{\mathbf{v}_j\}_{j=1}^J$ test locations.

Linear-time 2-sample test [Chwialkowski et al., 2015]

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}.$$

- Idea: change this to

$$\rho(\mathbb{P},\mathbb{Q}) := \rho\left(\mathbb{P},\mathbb{Q};\{\mathbf{v}_j\}_{j=1}^J\right) := \sqrt{\frac{1}{J}\sum_{j=1}^J [\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}$$

with random $\{\mathbf{v}_j\}_{j=1}^J$ test locations.

Is ρ a random metric? How do we estimate it? Distribution under H_0?

What is a random metric?

In short

It is a metric almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) = \rho(\mathbb{Q},\mathbb{P})$ almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) = \rho(\mathbb{Q},\mathbb{P})$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) \leqslant \rho(\mathbb{P},\mathbb{D}) + \rho(\mathbb{D},\mathbb{Q})$ almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) = \rho(\mathbb{Q},\mathbb{P})$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) \leqslant \rho(\mathbb{P},\mathbb{D}) + \rho(\mathbb{D},\mathbb{Q})$ almost surely.

$\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^{J} \subset \mathbb{R}^d$: reason of randomness.

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,
- *analytic:* $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$ *is analytic for any* $\mathbf{y} \in \mathbb{R}^d$.

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,
- *analytic:* $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$ *is analytic for any* $\mathbf{y} \in \mathbb{R}^d$.
- *characteristic:* μ_k *is injective,*

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,
- *analytic:* $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$ *is analytic for any* $\mathbf{y} \in \mathbb{R}^d$.
- *characteristic:* μ_k *is injective,*

then

$$\rho(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}$$

is a metric a.s. w.r.t. $\{\mathbf{v}_j\}_{j=1}^{J}$.

Why do analytic features work? – proof idea

- μ is injective and maps to analytic functions:
  - k: bounded, analytic $\Rightarrow$ elements of $\mathcal{H}_k$: analytic.
  - k: characteristic, bounded $\Rightarrow \mu = \mu_k$: well-defined, injective.

Why do analytic features work? – proof idea

- μ is injective and maps to analytic functions:
  - k: bounded, analytic $\Rightarrow$ elements of $\mathcal{H}_k$: analytic.
  - k: characteristic, bounded $\Rightarrow$ $\mu = \mu_k$: well-defined, injective.
- μ: characteristic $\Rightarrow$ for $\mathbb{P} \neq \mathbb{Q}$, $f := \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \neq 0$.

Why do analytic features work? – proof idea

- μ is injective and maps to analytic functions:
  - k: bounded, analytic $\Rightarrow$ elements of $\mathcal{H}_k$: analytic.
  - k: characteristic, bounded $\Rightarrow \mu = \mu_k$: well-defined, injective.
- μ: characteristic $\Rightarrow$ for $\mathbb{P} \neq \mathbb{Q}$, $f := \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \neq 0$.
- f: analytic, thus

$$\rho(\mathbb{P}, \mathbb{Q}) = \sqrt{\sum_{j=1}^{J} \left[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)\right]^2}$$

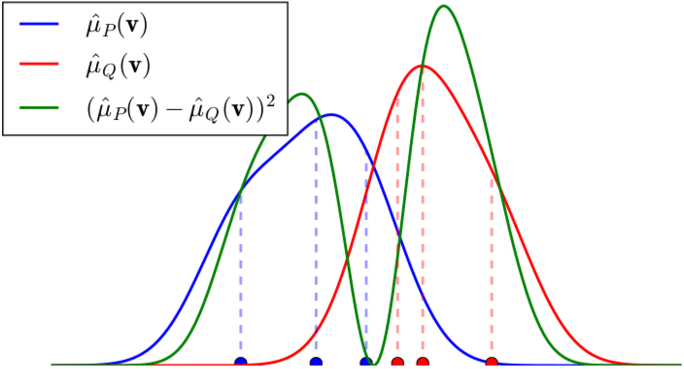
is a metric, a.s. w.r.t. $(\mathbf{v}_j \overset{i.i.d.}{\sim}) \; m \ll \lambda$. Reason: for an analytic $f \neq 0$, $m\{\mathbf{v} : f(\mathbf{v}) = 0\} = 0$.

Estimation

Compute

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2,$$

where $\hat{\mu}_{\mathbb{P}}(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i,\mathbf{v})$. Example using $k(\mathbf{x},\mathbf{v}) = e^{-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}}$:



Estimation – continued

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2$$

Estimation – continued

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2$$

$$= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_i,\mathbf{v}_j) - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{y}_i,\mathbf{v}_j)\right]^2$$

Estimation – continued

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2$$

$$= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_i,\mathbf{v}_j) - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{y}_i,\mathbf{v}_j)\right]^2 = \frac{1}{J}\sum_{j=1}^{J}(\bar{\mathbf{z}}_n)_j^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

where $\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}\underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^{J}}_{=:\mathbf{z}_i} \in \mathbb{R}^J$.

Estimation – continued

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2$$

$$= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_i,\mathbf{v}_j) - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{y}_i,\mathbf{v}_j)\right]^2 = \frac{1}{J}\sum_{j=1}^{J}(\bar{\mathbf{z}}_n)_j^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

where $\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}\underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^{J}}_{=:\mathbf{z}_i} \in \mathbb{R}^J$.

- Good news: estimation is linear in n!
- Bad news: intractable null distr. $= \sqrt{n}\widehat{\rho^2}(\mathbb{P},\mathbb{P}) \xrightarrow{d}$ sum of J correlated χ^2.

Normalized version gives tractable (asymptotic) null

- Modified test statistic:

$$\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T \mathbf{\Sigma}_n^{-1} \bar{\mathbf{z}}_n,$$

  where $\mathbf{\Sigma}_n = cov\left(\{\mathbf{z}_i\}_{i=1}^n\right)$.
- Under H_0:
  - $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$. $\Rightarrow$ Easy to get the $(1-\alpha)$-quantile!

Notes

- Characteristic functions – 'poor' choice:

$$\rho_2(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\phi_{\mathbb{P}}(\mathbf{v}_j) - \phi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

[Chwialkowski et al., 2015, Prop. 1]: It fails to distinguish a large class of distributions.

Notes

- Characteristic functions – 'poor' choice:

$$\rho_2(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\phi_{\mathbb{P}}(\mathbf{v}_j) - \phi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

  [Chwialkowski et al., 2015, Prop. 1]: It fails to distinguish a large class of distributions.

- [Moulines et al., 2007]:

$$\rho_3(\mathbb{P},\mathbb{Q}) := \frac{n_x n_y}{n}\left\|C^{-\frac{1}{2}}(\mu_{\mathbb{Q}} - \mu_{\mathbb{P}})\right\|_{\mathcal{H}_k},$$

$$C = \frac{n_x}{n_x + n_y}C_{xx} + \frac{n_y}{n_x + n_y}C_{yy} : \text{ pooled covariance operator.}$$

Notes

- Characteristic functions – 'poor' choice:

$$\rho_2(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\phi_{\mathbb{P}}(\mathbf{v}_j) - \phi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

  [Chwialkowski et al., 2015, Prop. 1]: It fails to distinguish a large class of distributions.

- [Moulines et al., 2007]:

$$\rho_3(\mathbb{P},\mathbb{Q}) := \frac{n_x n_y}{n}\left\| C^{-\frac{1}{2}}(\mu_{\mathbb{Q}} - \mu_{\mathbb{P}})\right\|_{\mathcal{H}_k},$$

$$C = \frac{n_x}{n_x + n_y} C_{xx} + \frac{n_y}{n_x + n_y} C_{yy} : \text{ pooled covariance operator.}$$

  Computational cost: high (cubic).

Notes – continued

- Until now: spatial domain.
- Smoothed characteristic functions:

$$\psi_{\mathbb{P}}(\mathbf{t}) = \int_{\mathbb{R}^d} \phi_{\mathbb{P}}(\boldsymbol{\omega})\ell(\mathbf{t}-\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}, \quad \mathbf{t} \in \mathbb{R}^d,$$

$$\rho_4(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\psi_{\mathbb{P}}(\mathbf{v}_j) - \psi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

Notes – continued

- Until now: spatial domain.
- Smoothed characteristic functions:

$$\psi_{\mathbb{P}}(\mathbf{t}) = \int_{\mathbb{R}^d} \phi_{\mathbb{P}}(\boldsymbol{\omega})\ell(\mathbf{t}-\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}, \quad \mathbf{t} \in \mathbb{R}^d,$$

$$\rho_4(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\psi_{\mathbb{P}}(\mathbf{v}_j) - \psi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

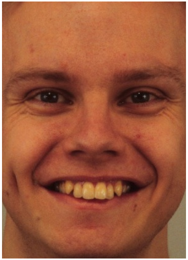
- Notes:
  - For analytic smoothing kernels (ℓ), it works.
  - It is more sensitive to differences in the frequency domain.

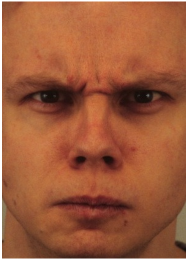
Linear-time high-power two-sample testing

Example-1: NLP

- Given: two categories of documents (Bayesian inference, neuroscience).
- Task:
  - test their distinguishability,
  - most discriminative words $\rightarrow$ interpretability.

Example-2: computer vision





- Given: two sets of faces (happy, angry).
- Task:
  - check if they are different,
  - determine the most discriminative features/regions.

One-page summary [Jitkrittum et al., 2016a]

- We get a nonparametric t-test.
- It gives a reason why H_0 is rejected.
- It is
  - adaptive → high test power.
  - fast (linear time).

One-page summary [Jitkrittum et al., 2016a]

- We get a nonparametric t-test.
- It gives a reason why H_0 is rejected.
- It is
  - adaptive $\rightarrow$ high test power.
  - fast (linear time).

Code:

- `https://github.com/wittawatj/interpretable-test`

Idea

- Until this point: test locations ($\mathcal{V}$) are fixed.
- Instead: choose $\theta = \{\mathcal{V}, \sigma\}$ to

  maximize lower bound on the test power.

Idea

- Until this point: test locations ($\mathcal{V}$) are fixed.
- Instead: choose $\theta = \{\mathcal{V}, \sigma\}$ to

  maximize lower bound on the test power.

Theorem (Lower bound on power, for large n)

Test power $\geqslant L(\lambda_n)$; L: explicit function, monotonically increasing.

- Here,
  - $\lambda_n = n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$: population version of $\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T\boldsymbol{\Sigma}_n^{-1}\bar{\mathbf{z}}_n$.
  - $\boldsymbol{\mu} = \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1]$, $\boldsymbol{\Sigma} = \mathbb{E}_{\mathbf{xy}}\left[(\mathbf{z}_1 - \boldsymbol{\mu})(\mathbf{z}_1 - \boldsymbol{\mu})^T\right]$.

Convergence of the λ_n estimator

But λ_n is unknown. Split (X, Y) into (X_{tr}, Y_{tr}) and (X_{te}, Y_{te}).

- Locations, kernel parameter: $\hat{\theta} = \arg\max_{\theta} \hat{\lambda}^{tr}_{\frac{n}{2}}(\theta)$.

Convergence of the λ_n estimator

But λ_n is unknown. Split (X, Y) into (X_{tr}, Y_{tr}) and (X_{te}, Y_{te}).

- Locations, kernel parameter: $\hat{\theta} = \arg\max_{\theta} \hat{\lambda}^{tr}_{\frac{n}{2}}(\theta)$.
- Test statistic: $\hat{\lambda}^{te}_{\frac{n}{2}}(\hat{\theta})$.

Convergence of the λ_n estimator

Theorem (Guarantee on objective approximation, $\gamma_n \to 0$)

$$\sup_{\mathcal{V},\mathcal{K}} \left|\bar{\mathbf{z}}_n^T (\boldsymbol{\Sigma}_n + \gamma_n)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right| = \mathcal{O}\left(n^{-\frac{1}{4}}\right).$$

Convergence of the λ_n estimator

Theorem (Guarantee on objective approximation, $\gamma_n \to 0$)

$$\sup_{\mathcal{V},\mathcal{K}} \left|\bar{\mathbf{z}}_n^T (\mathbf{\Sigma}_n + \gamma_n)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}\right| = \mathcal{O}\left(n^{-\frac{1}{4}}\right).$$

Examples:

$$\mathcal{K} = \left\{ k_\sigma(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} : \sigma > 0 \right\},$$

$$\mathcal{K} = \left\{ k_{\mathbf{A}}(\mathbf{x},\mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \mathbf{A}(\mathbf{x}-\mathbf{y})} : \mathbf{A} \succ 0 \right\}.$$

Proof idea

- Lower bound on the test power:
  - $|\hat{\lambda}_n - \lambda_n| \lesssim \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_F$.
  - Bound the r.h.s. by Hoeffding inequality $\Rightarrow P(|\hat{\lambda}_n - \lambda_n| \geqslant t)$.
  - By reparameterization: $P(\hat{\lambda}_n \geqslant T_\alpha)$ bound.

Proof idea

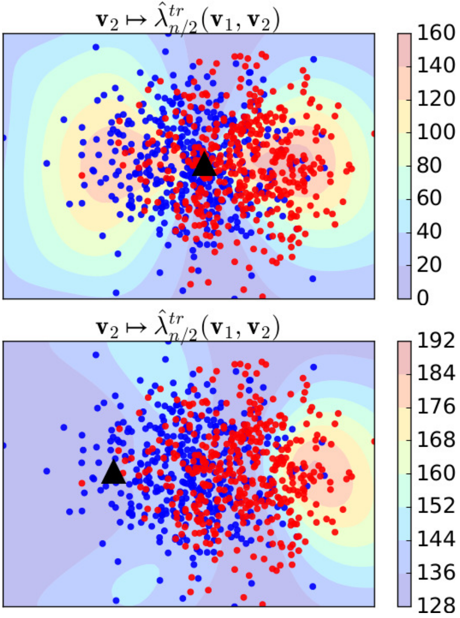
- Lower bound on the test power:
  - $|\hat{\lambda}_n - \lambda_n| \lesssim \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_F$.
  - Bound the r.h.s. by Hoeffding inequality $\Rightarrow P(|\hat{\lambda}_n - \lambda_n| \geqslant t)$.
  - By reparameterization: $P(\hat{\lambda}_n \geqslant T_\alpha)$ bound.
- Uniformly $\hat{\lambda}_n \approx \lambda_n$:
  - Reduction to bounding $\sup_{\mathcal{V},\mathcal{S}} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$, $\sup_{\mathcal{V},\mathcal{S}} \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_F$.
  - Empirical processes, Dudley entropy bound.

Non-convexity, informative features

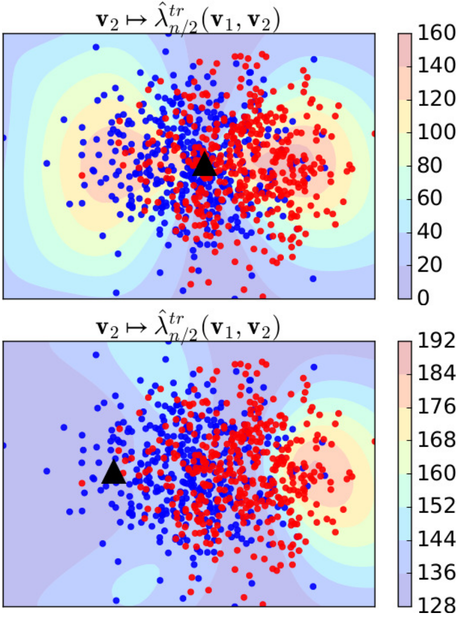
- 2D problem:

$$\mathbb{P} := \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbb{Q} := \mathcal{N}(\mathbf{e}_1, \mathbf{I}).$$

- $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$. Fix $\mathbf{v}_1$ to the triangle.
- $\mathbf{v}_2 \mapsto \hat{\lambda}_n(\{\mathbf{v}_1, \mathbf{v}_2\})$: contour plot.



Non-convexity, informative features



- Nearby locations: do not increase discrimininability.
- Non-convexity: reveals multiple ways to capture the difference.

Computational complexity

- Optimization & testing: linear in n.
- Testing: $\mathcal{O}\left(ndJ + nJ^2 + J^3\right)$.
- Optimization: $\mathcal{O}\left(ndJ^2 + J^3\right)$ per gradient ascent.

Number of locations (J)

- Small J:
  - often enough to detect the difference of $\mathbb{P}$ & $\mathbb{Q}$.
  - few distinguishing regions to reject H_0.
  - faster test.

Number of locations (J)

- Very large J:
  - test power need not increase monotonically in J (more locations $\Rightarrow$ statistic can gain in variance).
  - defeats the purpose of a linear-time test.

Numerical demos

Parameter settings

- Gaussian kernel (σ). $\alpha = 0.01$. $J = 1$. Repeat 500 trials.
- Report

$$P(\text{reject } H_0) \approx \frac{\#\text{times } \hat{\lambda}_n > T_\alpha \text{ holds}}{\#\text{trials}}.$$

- Compare 4 methods
  - **ME-full**: Optimize $\mathcal{V}$ and Gaussian bandwidth σ.
  - **ME-grid**: Optimize σ. Random $\mathcal{V}$ [Chwialkowski et al., 2015].
  - **MMD-quad**: Test with quadratic-time MMD [Gretton et al., 2012].
  - **MMD-lin**: Test with linear-time MMD [Gretton et al., 2012].
- Optimize kernels to power in MMD-lin, MMD-quad.

NLP: discrimination of document categories

- 5903 NIPS papers (1988-2015).
- Keyword-based category assignment into 4 groups:
  - Bayesian inference, Deep learning, Learning theory, Neuroscience
- $d = 2000$ nouns. TF-IDF representation.

Problem	n^{te}	**ME-full**	ME-grid	MMD-quad	MMD-lin
1. Bayes-Bayes	215	.012	.018	.022	.008
2. Bayes-Deep	216	.954	.034	.906	.262
3. Bayes-Learn	138	.990	.774	1.00	.238
4. Bayes-Neuro	394	1.00	.300	.952	.972
5. Learn-Deep	149	.956	.052	.876	.500
6. Learn-Neuro	146	.960	.572	1.00	.538

- Performance of ME-full $[\mathcal{O}(n)]$ is comparable to MMD-quad $[\mathcal{O}(n^2)]$.

NLP: most/least discriminative words

- Aggregating over trials; example: 'Bayes-Neuro'.
- Most discriminative words:

  spike, markov, cortex, dropout, recurr, iii, gibb.

  - learned test locations: highly interpretable,
  - 'markov', 'gibb' ($\Leftarrow$ Gibbs): Bayesian inference,
  - 'spike', 'cortex': key terms in neuroscience.

NLP: most/least discriminative words

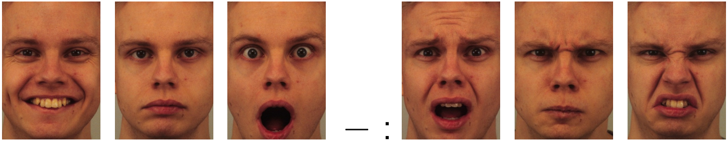
- Aggregating over trials; example: 'Bayes-Neuro'.

- Least dicriminative ones:

  circumfer, bra, dominiqu, rhino, mitra, kid, impostor.

Distinguish positive/negative emotions

- Karolinska Directed Emotional Faces (KDEF) [Lundqvist et al., 1998].
- 70 actors = 35 females and 35 males.
- $d = 48 \times 34 = 1632$. Grayscale. Pixel features.



Problem	n^{te}	**ME-full**	ME-grid	MMD-quad	MMD-lin
$\pm$ vs. $\pm$	201	.010	.012	.018	.008
$+$ vs. $-$	201	.998	.656	1.00	.578

- Learned test location (averaged) = 

Linear-time high-power two-sample testing: finished

Linear-time high-power independence testing

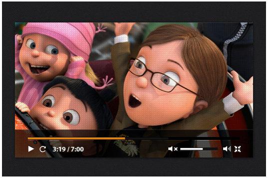
Example: dependency testing of media annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

Example: dependency testing of media annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

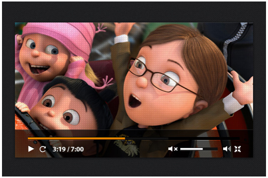
  - (video, caption) pairs



Example: dependency testing of media annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

  - (video, caption) pairs



- $\{(x_i, y_i)\}_{i=1}^n \xrightarrow{?} H_0 : \mathbb{P}_{xy} = \mathbb{P}_x\mathbb{P}_y, \ H_1 : \mathbb{P}_{xy} \neq \mathbb{P}_x\mathbb{P}_y.$

2-sample test → independence test

Until now:

- adaptive linear-time 2-sample test (automatic parameter tuning).

2-sample test → independence test

2-sample test:

$$MMD(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}, \quad \rho(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2},$$

2-sample test → independence test

2-sample test:

$$MMD(\mathbb{P},\mathbb{Q}) = \left\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\|_{\mathcal{H}_k}, \quad \rho(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2},$$

Independence test [Jitkrittum et al., 2016b]:

$$HSIC(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}, \quad FSIC(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j, \mathbf{w}_j)}$$

2-sample test $\to$ independence test

2-sample test:

$$MMD(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}, \quad \rho(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2},$$

Independence test [Jitkrittum et al., 2016b]:

$$HSIC(x,y) = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}, \quad FSIC(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)},$$

with $u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w})$ witness function.

FSIC: covariance view

By rewriting

$$\begin{aligned} u(\mathbf{v},\mathbf{w}) &= \mu_{\mathbf{xy}}(\mathbf{v},\mathbf{w}) - \mu_{\mathbf{x}}(\mathbf{v})\mu_{\mathbf{y}}(\mathbf{w}) \\ &= \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x},\mathbf{v})\ell(\mathbf{y},\mathbf{w})] - \mathbb{E}_{\mathbf{x}}[k(\mathbf{x},\mathbf{v})]\mathbb{E}_{\mathbf{y}}[\ell(\mathbf{y},\mathbf{w})] \\ &= cov_{\mathbf{xy}}\left(k(\mathbf{x},\mathbf{v}),\ell(\mathbf{y},\mathbf{w})\right). \end{aligned}$$

$\Rightarrow$ We picked the $(\mathbf{v},\mathbf{w})^{th}$ entry of

$$\begin{aligned} C^c_{\mathbf{xy}} &= \mathbb{E}_{xy}\left[\varphi(\mathbf{x})\otimes\psi(\mathbf{y})\right] - \mu_{\mathbf{x}}\otimes\mu_{\mathbf{y}}, \\ HSIC &= \left\|C^c_{\mathbf{xy}}\right\|_{HS}. \end{aligned}$$

FSIC is an independence measure

Theorem

If $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *are bounded, characteristic, analytic kernels* [$\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$]*, then almost surely*

$$FSIC(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} \perp \mathbf{y}.$$

FSIC is an independence measure

Theorem

If $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *are bounded, characteristic, analytic kernels* [$\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$]*, then almost surely*

$$FSIC(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} \perp \mathbf{y}.$$

Consequence

FSIC can be applied in ISA, feature selection, outlier-robust image registration, . . .

Empirical estimator for FSIC

$$FSIC^2(\mathbf{x},\mathbf{y}) = \frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j), \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$\widehat{FSIC}^2(\mathbf{x},\mathbf{y}) = \frac{1}{J}\sum_{j=1}^{J} \hat{u}^2(\mathbf{v}_j,\mathbf{w}_j), \quad \hat{u}(\mathbf{v},\mathbf{w}) = \widehat{\mu_{xy}}(\mathbf{v},\mathbf{w}) - (\widehat{\mu_x\mu_y})(\mathbf{v},\mathbf{w}),$$

$$= \frac{1}{J}\|\mathbf{u}\|_2^2$$

Empirical estimator for FSIC

$$FSIC^2(\mathbf{x},\mathbf{y}) = \frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j), \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$\widehat{FSIC}^2(\mathbf{x},\mathbf{y}) = \frac{1}{J}\sum_{j=1}^{J} \hat{u}^2(\mathbf{v}_j,\mathbf{w}_j), \quad \hat{u}(\mathbf{v},\mathbf{w}) = \widehat{\mu_{xy}}(\mathbf{v},\mathbf{w}) - (\widehat{\mu_x\mu_y})(\mathbf{v},\mathbf{w}),$$

$$= \frac{1}{J}\|\mathbf{u}\|_2^2,$$

where we use the unbiased estimators [2nd = '$\mu_x(\mathbf{v})\mu_y(\mathbf{w})$ - diag']:

$$\widehat{\mu_{xy}}(\mathbf{v},\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i,\mathbf{v})\ell(\mathbf{y}_i,\mathbf{w}),$$

$$\widehat{\mu_x\mu_y}(\mathbf{v},\mathbf{w}) = \frac{1}{n(n-1)}\sum_{i\neq j} k(\mathbf{x}_i,\mathbf{v})\ell(\mathbf{y}_j,\mathbf{w}).$$

Asymptotic distribution of $\hat{\mathbf{u}}$

For fixed $(\mathbf{v}, \mathbf{w})$:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \frac{2}{n(n-1)} \sum_{i<j} h_{\mathbf{v},\mathbf{w}} \left((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j) \right),$$

$$h_{\mathbf{v},\mathbf{w}} \left((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \right) = \frac{1}{2} \left[k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v}) \right] \left[\ell(\mathbf{y}, \mathbf{w}) - \ell(\mathbf{y}', \mathbf{w}) \right]$$

Asymptotic distribution of $\hat{\mathbf{u}}$

For fixed $(\mathbf{v}, \mathbf{w})$:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \frac{2}{n(n-1)} \sum_{i<j} h_{\mathbf{v},\mathbf{w}}\left((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)\right),$$

$$h_{\mathbf{v},\mathbf{w}}\left((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')\right) = \frac{1}{2}\left[k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v})\right]\left[\ell(\mathbf{y}, \mathbf{w}) - \ell(\mathbf{y}', \mathbf{w})\right],$$

thus $\xrightarrow{\text{theory of U-statistics}}$

Theorem (Asymptotic normality)

For any fixed locations $\mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^J$, $\hat{\mathbf{u}} := [\hat{u}(\mathbf{v}_j, \mathbf{w}_j)]_{j=1}^J$

$$\sqrt{n}\left(\hat{\mathbf{u}} - \mathbf{u}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}),$$

$$\Sigma_{ij} = cov_{\mathbf{xy}}\left(\hat{u}(\mathbf{v}_i, \mathbf{w}_i), \hat{u}(\mathbf{v}_j, \mathbf{w}_j)\right).$$

NFSIC = FSIC + whitening

- $n\widehat{FSIC}^2(x,y) = n\frac{\|\mathbf{u}\|_2^2}{J}$: asymptotically sum of correlated χ^2-s.

NFSIC = FSIC + whitening

- $n\widehat{FSIC}^2(x,y) = n\frac{\|\mathbf{u}\|_2^2}{J}$: asymptotically sum of correlated χ^2-s.
- Quantile: hard. $\Rightarrow$ With the whitening trick:

Theorem

- *Under H_0: with $\gamma_n \to 0$*

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T\left(\hat{\mathbf{\Sigma}}_n + \gamma_n \mathbf{I}_J\right)^{-1}\hat{\mathbf{u}} \xrightarrow{d} \chi^2(J).$$

NFSIC = FSIC + whitening

- $n\widehat{FSIC}^2(x,y) = n\frac{\|\mathbf{u}\|_2^2}{J}$: asymptotically sum of correlated χ^2-s.
- Quantile: hard. $\Rightarrow$ With the whitening trick:

Theorem

- *Under H_0: with $\gamma_n \to 0$*

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\mathbf{\Sigma}}_n + \gamma_n \mathbf{I}_J\right)^{-1} \hat{\mathbf{u}} \xrightarrow{d} \chi^2(J).$$

- *Under H_1: we get a consistent test (i.e., power $\to 1$).*

NFSIC can be estimated easily

Test statistic:

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\mathbf{\Sigma}}_n + \gamma_n \mathbf{I}_J\right)^{-1} \hat{\mathbf{u}}.$$

Estimator: no $n \times n$ Gram matrix

- $\mathbf{K} := [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J \times n}$, $\mathbf{L} := [\ell(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J \times n}$,
- $\hat{\mathbf{\Sigma}}_n = \frac{\mathbf{\Gamma}\mathbf{\Gamma}^T}{n}$, $\mathbf{\Gamma} = (\mathbf{K}\mathbf{H}_n) \circ (\mathbf{L}\mathbf{H}_n) - \hat{\mathbf{u}}\mathbf{1}_n^T$, $\hat{\mathbf{u}} := \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.

Computational time:

$$\mathcal{O}\left(J^3 + J^2 n + (d_x + d_y) J n\right).$$

NFSIC can be estimated easily

Test statistic:

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\boldsymbol{\Sigma}}_n + \gamma_n \mathbf{I}_J\right)^{-1} \hat{\mathbf{u}}.$$

Estimator: no $n \times n$ Gram matrix

- $\mathbf{K} := [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J\times n}$, $\mathbf{L} := [\ell(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J\times n}$,
- $\hat{\boldsymbol{\Sigma}}_n = \frac{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T}{n}$, $\boldsymbol{\Gamma} = (\mathbf{K}\mathbf{H}_n) \circ (\mathbf{L}\mathbf{H}_n) - \hat{\mathbf{u}}\mathbf{1}_n^T$, $\hat{\mathbf{u}} := \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.

Computational time:

$$\mathcal{O}\left(J^3 + J^2 n + (d_x + d_y) J n\right).$$

Code with demos:

https://github.com/wittawatj/fsic-test

Choosing the locations & kernel parameters

- Consistent test: for $\forall \mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j\}_{j=1}^{J}$ and kernel parameters.

Choosing the locations & kernel parameters

- Consistent test: for $\forall \mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j\}_{j=1}^J$ and kernel parameters.
- Choose the power proxy maximizer.

Theorem

Let $NFSIC^2(x, y) = \lambda_n = n\mathbf{u}^T \mathbf{\Sigma}^{-1}\mathbf{u}$*. For large* n*,*

$$\text{test power} \geqslant L(\lambda_n),$$

L*: monotonically increasing.*

Choosing the locations & kernel parameters

- Consistent test: for $\forall \mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j\}_{j=1}^J$ and kernel parameters.
- Choose the power proxy maximizer.

Theorem

Let $NFSIC^2(x, y) = \lambda_n = n\mathbf{u}^T\mathbf{\Sigma}^{-1}\mathbf{u}$. *For large* n,

$$\text{test power} \geqslant L(\lambda_n),$$

L*: monotonically increasing.*

- In practice: data-splitting (a la 2-sample testing).

HSIC versus FSIC

Question

Which one to choose?

- *HSIC* $= \|u\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.
- *FSIC* $= \|u\|_{L^2(\mathcal{V})}$, $\mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^J$.

HSIC versus FSIC

Question

Which one to choose?

- *HSIC* $= \|u\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.
  - When $p_{xy} - p_x p_y$ is diffuse, close to flat.
- *FSIC* $= \|u\|_{L^2(\mathcal{V})}$, $\mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^J$.

HSIC versus FSIC

Question

Which one to choose?

- *HSIC* $= \|u\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.
  - When $p_{xy} - p_x p_y$ is diffuse, close to flat.
- *FSIC* $= \|u\|_{L^2(\mathcal{V})}$, $\mathcal{V} = \{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^J$.
  - When $p_{xy} - p_x p_y$ is local, with many peaks.

Demo settings

- k, ℓ: Gaussian. $J = 10$.
- Report: rejection rate of H_0.
- Compare 6 methods:

Method	Description	Tuning	Test size	Complexity
NFSIC-opt	Studied	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heuristic	n	$\mathcal{O}(n^2)$
NyHSIC	Nyström + HSIC	Median heuristic	n	$\mathcal{O}(n)$
FHSIC	RFF + HSIC	Median heuristic	n	$\mathcal{O}(n)$
RDC	RFF + CCA	Median heuristic	n	$\mathcal{O}(n \log n)$

Demo-1: million song data

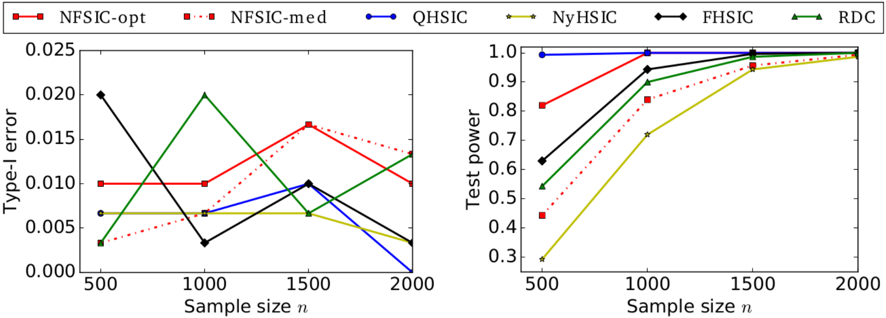
Song ($\mathbf{x}$) vs. year of release (y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $\mathbf{x} \in \mathbb{R}^{90=d_x}$: audio features.
- Left: break (x, y) pairs, i.e. H_0 holds; right: H_1 is true.

Demo-1: million song data

Song ($\mathbf{x}$) vs. year of release (y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $\mathbf{x} \in \mathbb{R}^{90=d_x}$: audio features.
- Left: break (x, y) pairs, i.e. H_0 holds; right: H_1 is true.



Demo-2: videos and captions

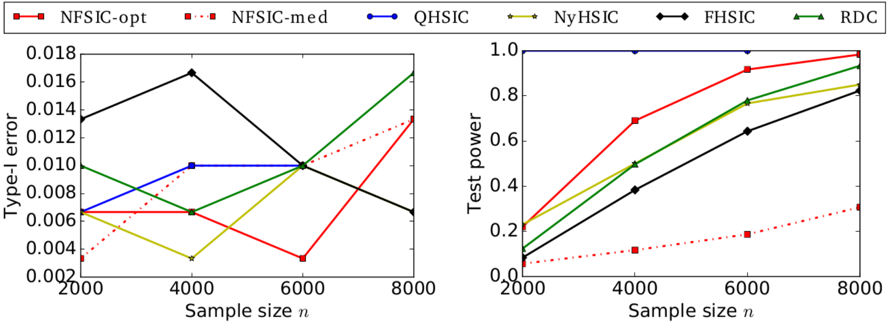
Youtube video ($\mathbf{x}$) vs. caption ($\mathbf{y}$).

- VideoStory46K [Habibian et al., 2014]
- $\mathbf{x} \in \mathbb{R}^{2000=d_x}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $\mathbf{y} \in \mathbb{R}^{1878=d_y}$: bag of words. TF.
- Left: break (x, y) pairs, i.e. H_0 holds; right: H_1 is true.

Demo-2: videos and captions

Youtube video ($\mathbf{x}$) vs. caption ($\mathbf{y}$).

- VideoStory46K [Habibian et al., 2014]
- $\mathbf{x} \in \mathbb{R}^{2000=d_x}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $\mathbf{y} \in \mathbb{R}^{1878=d_y}$: bag of words. TF.
- Left: break (x, y) pairs, i.e. H_0 holds; right: H_1 is true.



Linear-time goodness-of-fit testing: Stein operator & analytical kernels [Jitkrittum et al., 2017]

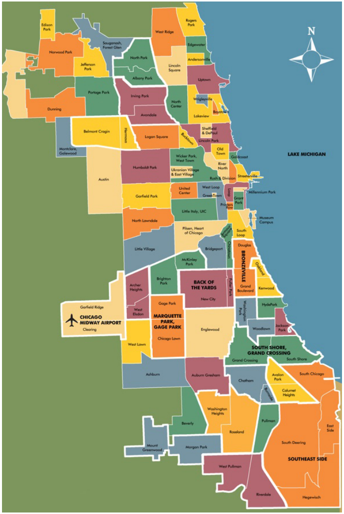
Given:

- Density/model: p.

Linear-time goodness-of-fit testing: Stein operator & analytical kernels [Jitkrittum et al., 2017]

Given:

- Density/model: p.
- Samples: $X = \{x_i\}_{i=1}^n \sim q$ (unknown).



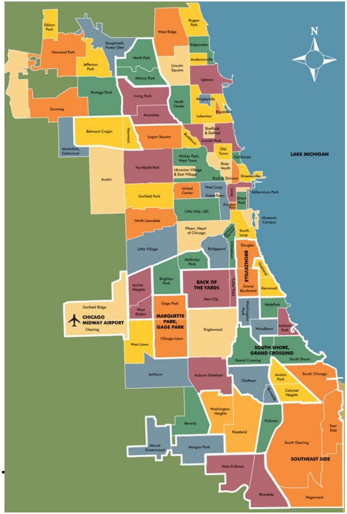
Linear-time goodness-of-fit testing: Stein operator & analytical kernels [Jitkrittum et al., 2017]

Given:

- Density/model: p.
- Samples: $X = \{x_i\}_{i=1}^n \sim q$ (unknown).

Problem: using p, X test

$$H_0 : p = q, \quad \text{vs}$$
$$H_1 : p \neq q.$$



Quick summary:

- Best paper award (NIPS-2017, 3/3240).
- Demo: criminal data analysis.
- Code: `https://github.com/wittawatj/kernel-gof`

Summary

- Dependency measures, distances: KCCA, HSIC, MMD.
- Mean embedding, cross-covariance operator.

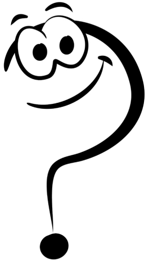
Summary

- Dependency measures, distances: KCCA, HSIC, MMD.
- Mean embedding, cross-covariance operator.
- Applications:
  - ISA, distribution regression, image registration, feature selection,
  - hypothesis testing.

Summary

- Dependency measures, distances: KCCA, HSIC, MMD.
- Mean embedding, cross-covariance operator.
- Applications:
    - ISA, distribution regression, image registration, feature selection,
    - hypothesis testing.
- Hypothesis testing:
    - quadratic methods,
    - scaling: block-variants, Nyström, RFF,
    - linear-time adaptive nonparametric tests.

Thank you for the attention!



Altun, Y. and Smola, A. (2006).
Unifying divergence minimization and statistical inference via convex duality.
In *Conference on Learning Theory (COLT)*, pages 139–153.

Bach, F. R. and Jordan, M. I. (2002).
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48.

Baker, C. R. (1973).
Joint measures and cross-covariance operators.
Transactions of the American Mathematical Society, 186:273–289.

Baringhaus, L. and Franz, C. (2004).
On a new multivariate two-sample test.
Journal of Multivariate Analysis, 88:190–206.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984).
Harmonic Analysis on Semigroups.
Springer-Verlag.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011).
The million song dataset.
In *International Conference on Music Information Retrieval (ISMIR)*.

Blanchard, G., Lee, G., and Scott, C. (2011).
Generalizing from several related classification tasks to a new unlabeled sample.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186.

Brault, R., Heinonen, M., and d'Alché-Buc, F. (2016).
Random Fourier features for operator-valued kernels.

In *Asian Conference in Machine Learning (ACML; JMLR W&CP)*, volume 63, pages 110–125.

Caponnetto, A. and De Vito, E. (2007).
Optimal rates for regularized least-squares algorithm.
Foundations of Computational Mathematics, 7:331–368.

Cardoso, J.-F. (1998).
Multidimensional independent component analysis.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1941–1944.

Carmeli, C., Vito, E. D., Toigo, A., and Umanitá, V. (2010).
Vector valued reproducing kernel Hilbert spaces and universality.
Analysis and Applications, 8:19–61.

Christmann, A. and Steinwart, I. (2010).
Universal kernels on non-standard input spaces.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 406–414.

Chwialkowski, K. and Gretton, A. (2014).
A kernel independence test for random processes.
In *International Conference on Machine Learning (ICML; JMLR W&CP)*, volume 32, page 14221430.

Chwialkowski, K., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015).
Fast two-sample testing with analytic representations of probability measures.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1972–1980.

Chwialkowski, K., Sejdinovic, D., and Gretton, A. (2014).
A wild bootstrap for degenerate kernel tests.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 3608–3616.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.

In *International Conference on Machine Learning (ICML)*, pages 2606–2615.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 625–632.

Csörgo, S. and Totik, V. (1983).
On how long interval is the empirical characteristic function uniformly consistent?
Acta Scientiarum Mathematicarum, 45:141–149.

Cuturi, M. (2011).
Fast global alignment kernels.
In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

Diestel, J. and Uhl, J. J. (1977).
Vector Measures.
American Mathematical Society. Providence.

Dinculeanu, N. (2000).
Vector Integration and Stochastic Integration in Banach Spaces.
Wiley.

Drineas, P. and Mahoney, M. W. (2005).
On the Nyström method for approximating a Gram matrix for improved kernel-based learning.
Journal of Machine Learning Research, 6:2153–2175.

Dudley, R. M. (2004).
Real Analysis and Probability.
Cambridge University Press.

Fang, K.-T., Kotz, S., and Ng, K. W. (1990).
Symmetric multivariate and related distributions.

Chapman and Hall.

Fukumizu, K., Bach, F., and Jordan, M. (2004).
Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces.
Journal of Machine Learning Research, 5:73–99.

Fukumizu, K., Bach, F., and Jordan, M. (2009a).
Kernel dimension reduction in regression.
The Annals of Statistics, 37(4):1871–1905.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007).
Statistical consistency of kernel canonical correlation analysis.
Journal of Machine Learning Research, 8:361–383.

Fukumizu, K., Gretton, A., Schölkopf, B., and Sriperumbudur, B. K. (2009b).
Characteristic kernels on groups and semigroups.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 473–480.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gretton, A. (2015).
A simpler condition for consistency of a kernel independence test.
Technical report, University College London. (https://arxiv.org/abs/1501.06103).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.

Journal of Machine Learning Research, 13:723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a).
Measuring statistical dependence with Hilbert-Schmidt norms.
In *Algorithmic Learning Theory (ALT)*, pages 63–78.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009).
A fast, consistent kernel two-sample test.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 673–681.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007).
A kernel statistical test of independence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008).

A kernel statistical test of independence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b).
Kernel methods for measuring independence.
Journal of Machine Learning Research, 6:2075–2129.

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.
In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Habibian, A., Mensink, T., and Snoek, C. G. (2014).
Videostory: A new multimedia embedding for few-example recognition and translation of events.
In *ACM International Conference on Multimedia*, pages 17–26.

Haussler, D. (1999).
Convolution kernels on discrete structures.

Technical report, Department of Computer Science, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2016).
The Kendall and Mallows kernels for permutations.
In *International Conference on Machine Learning (ICML; PMLR)*, volume 37, pages 2982–2990.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016a).
Interpretable distribution features with maximum testing power.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 181–189.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2016b).
An adaptive test of independence with analytic kernel embeddings.
Technical report.
(https://arxiv.org/abs/1610.04782).

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017).
A linear-time kernel goodness-of-fit test.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 261–270.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994).

Continuous univariate distributions (volume 1).
John Wiley & Sons.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

Klebanov, L. (2005).
N-Distances and Their Applications.
Charles University, Prague.

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kybic, J. (2004).
High-dimensional mutual information estimation for image registration.

In *IEEE International Conference on Image Processing (ICIP)*, pages 1779–1782.

Lancaster, H. O. (1969).
The Chi-squared Distribution.
John Wiley and Sons Inc.

Leucht, A. and H.Neumann, M. (2013).
Dependent wild bootstrap for degenerate U- and V-statistics.
Journal of Multivariate Analysis, 117:257–280.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993).
Canonical correlation analysis when the data are curves.
Journal of the Royal Statistical Society, Series B (Methodological), 55(3):725–740.

Liu, Q., Lee, J., and Jordan, M. (2016).
A Kernelized Stein Discrepancy for Goodness-of-fit Tests.
In *International Conference on Machine Learning (ICML)*, pages 276–284.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Lopez-Paz, D. (2016).
From Dependence to Causation.
PhD thesis, University of Cambridge.

Lopez-Paz, D., Sra, S., Smola, A., Ghahramani, Z., and Schölkopf, B. (2014).
Randomized nonlinear component analysis.
In *International Conference on Machine Learning (ICML)*, pages 1359–1367.

Lundqvist, D., Flykt, A., and Öhman, A. (1998).
The Karolinska directed emotional faces-KDEF.
Technical report, ISBN 91-630-7164-9.

Lyons, R. (2013).

Distance covariance in metric spaces.
Annals of Probability, 41:3284–3305.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2009).
Nonextensive information theoretic kernels on measures.
The Journal of Machine Learning Research, 10:935–975.

Micchelli, C. A., Xu, Y., and Zhang, H. (2006).
Universal kernels.
Journal of Machine Learning Research, 7:2651–2667.

Moulines, É., Bach, F. R., and Harchaoui, Z. (2007).
Testing for homogenity with kernel Fisher discriminant analysis.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2011).
Learning from distributions via support measure machines.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017).
Kernel mean embedding of distributions: A review and beyond.
Foundations and Trends in Machine Learning, 10(1-2):1–141.

Müller, A. (1997).
Integral probability metrics and their generating classes of functions.
Advances in Applied Probability, 29:429–443.

Neemuchwala, H., Hero, A., Zabuawala, S., and Carson, P. (2007).
Image registration methods in high dimensional space.
International Journal of Imaging Systems and Technology, 16:130–145.

Nishiyama, Y. and Fukumizu, K. (2016).

Characteristic kernels and infinity divisibility.
Journal of Machine Learning Research, 17:1–28.

Peng, H., Long, F., and Ding, C. (2005).
Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8):1226–1238.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2017).
Kernel-based tests for joint independence.
Journal of the Royal Statistical Society: Series B (Statistical Methodology).

Póczos, B., Singh, A., Rinaldo, A., and Wasserman, L. (2013).
Distribution-free distribution regression.
In *International Conference on AI and Statistics (AISTATS; JMLR W&CP)*, volume 31, pages 507–515.

Póczos, B., Xiong, L., Sutherland, D., and Schneider, J. (2012).
Support distribution machines.
Technical report, Carnegie Mellon University. (http://arxiv.org/abs/1202.0302).

Rahimi, A. and Recht, B. (2007).
Random features for large-scale kernel machines.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184.

Reed, M. and Simon, B. (1980).
Methods of Modern Mathematical Physics, I: Functional Analysis.
Academic Press.

Rényi, A. (1959).
On measures of dependence.
Acta Mathematica Academiae Scientiarum Hungaricae, 10:441–451.

Rosasco, L., Santoro, M., Mosci, S., Verri, A., and Villa, S. (2010).
A regularization approach to nonlinear variable selection.
JMLR W&CP – International Conference on Artificial Intelligence and Statistics (AISTATS), 9:653–660.

Rosasco, L., Villa, S., Mosci, S., Santoro, M., and Verri, A. (2013).
Nonparametric sparsity and regularization.
Journal of Machine Learning Research, 14:1665–1714.

Rubenstein, P. K., Chwialkowski, K. P., and Gretton, A. (2016).
A kernel test for three-variable interactions with random processes.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 637–646.

Rudi, A. and Rosasco, L. (2017).
Generalization properties of learning with random features.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 3215–3225.

- Rudin, W. (1991).
Functional Analysis.
McGraw-Hill, USA.

- Sasvári, Z. (2013).
Multivariate Characteristic and Correlation Functions.
Walter de Gruyter.

- Sejdinovic, D., Gretton, A., and Bergsma, W. (2013a).
A kernel test for three-variable interactions.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132.

- Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013b).
Equivalence of distance-based and RKHS-based statistics in hypothesis testing.
Annals of Statistics, 41:2263–2291.

Serfling, R. J. (1980).
Approximation Theorems of Mathematical Statistics.
John Wiley & Sons.

Simon-Gabriel, C.-J. and Schölkopf, B. (2016).
Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions.
Technical report, MPI Tübingen.
https://arxiv.org/abs/1604.05251.

Snelson, E. and Ghahramani, Z. (2006).
Sparse Gaussian processes using pseudo-inputs.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1257–1264.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017).
Density estimation in infinite dimensional exponential families.
Journal of Machine Learning Research, 18(57):1–59.

- Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012).
  On the empirical estimation of integral probability metrics.
  Electronic Journal of Statistics, 6:1550–1599.

- Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2010a).
  On the relation between universality, characteristic kernels and rkhs embedding of measures.
  In *International Conference on AI and Statistics (AISTATS; JMLR W&CP)*, volume 9, pages 781–788.

- Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2011).
  Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint.
  In *Advances in Neural Information Processing Systems (NIPS)*, pages 1773–1781.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., and Lanckriet, G. R. G. (2010b).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Sriperumbudur, B. K. and Szabó, Z. (2015).
Optimal rates for random Fourier features.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1144–1152.

Steinwart, I. (2001).
On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 2:67–93.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.

Sutherland, D. J. and Schneider, J. (2015).
On the error of random Fourier features.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 862–871.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. (2015).
Two-stage sampled learning theory on distributions.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 948–957.

Szabó, Z., Póczos, B., and Lörincz, A. (2012).
Separation theorem for independent subspace analysis and its consequences.
Pattern Recognition, 45(4):1782–1791.

Szabó, Z. and Sriperumbudur, B. (2017).
Characteristic and universal tensor product kernels.
Technical report.
(`http://arxiv.org/abs/1708.08157`).

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.
Journal of Machine Learning Research, 17(152):1–40.

Székely, G. J. and Rizzo, M. L. (2004).
Testing for equal distributions in high dimension.
InterStat, 5.

Székely, G. J. and Rizzo, M. L. (2005).
A new test for multivariate normality.
Journal of Multivariate Analysis, 93:58–80.

Székely, G. J. and Rizzo, M. L. (2009).
Brownian distance covariance.
The Annals of Applied Statistics, 3:1236–1265.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007).
Measuring and testing dependence by correlation of distances.
The Annals of Statistics, 35:2769–2794.

Titsias, M. K. (2009).
Variational learning of inducing variables in sparse Gaussian processes.
Journal of Machine Learning Research, 5:567–574.

van de Geer, S. A. (2009).
Empirical Processes in M-Estimation.
Cambridge University Press.

van der Vaart, A. W. (1998).
Asymptotic Statistics.
Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996).
Weak Convergence and Empirical Processes.
Springer-Verlag.

Vishwanathan, S. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010).
Graph kernels.

Journal of Machine Learning Research, 11:1201–1242.

Waegeman, W., Pahikkala, T., Airola, A., Salakoski, T., Stock, M., and Baets, B. D. (2012).
A kernel-based framework for learning graded relations from data.
IEEE Transactions on Fuzzy Systems, 20:1090–1101.

Wang, H. and Schmid, C. (2013).
Action recognition with improved trajectories.
In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558.

Wendland, H. (2005).
Scattered Data Approximation.
Cambridge University Press.

Williams, C. K. I. and Seeger, M. (2001).
Using the Nyström method to speed up kernel machines.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 682–688.

Zaremba, W., Gretton, A., and Blaschko, M. (2013).
B-tests: Low variance kernel two-sample tests.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 755–763.

Zhang, H., Xu, Y., and Zhang, J. (2009).
Reproducing kernel Banach spaces for machine learning.
Journal of Machine Learning Research, 10:2741–2775.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2017).
Large-scale kernel methods for independence testing.
Statistics and Computing, pages 1–18.

Zhao, J. and Meng, D. (2015).
FastMMD: Ensemble of circular discrepancy for efficient two-sample test.
Neural Computation, 27:1345–1372.

Zinger, A. A., Kakosyan, A. V., and Klebanov, L. B. (1992).

A characterization of distributions by mean values of statistics and certain probabilistic metrics.
Journal of Soviet Mathematics.

Zolotarev, V. M. (1983).
Probability metrics.
Theory of Probability and its Applications, 28:278–302.