Kernel Methods, Divergence and Independence Measures, Hypothesis Testing

Zoltán Szabó – CMAP, École Polytechnique

Summer School on Data Science for Document Analysis and Understanding

INRIA, Palaiseau

July 15-17, 2019

Outline

- Applications:
  - Information theoretical objectives.
  - Testing.
- Classical information theory: $\mathbb{R}^d \xrightarrow{\text{diverse set of domains}}$
- Kernels, RKHS:
  - Linear $\rightarrow$ non-linear techniques.
  - Classification, regression, dimensionality reduction.
  - KCCA, MMD, HSIC.
- Hypothesis testing.

Information Theoretical Objectives

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

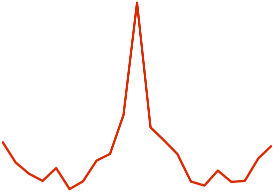
Given two images:


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: $\boldsymbol{\Theta}$.

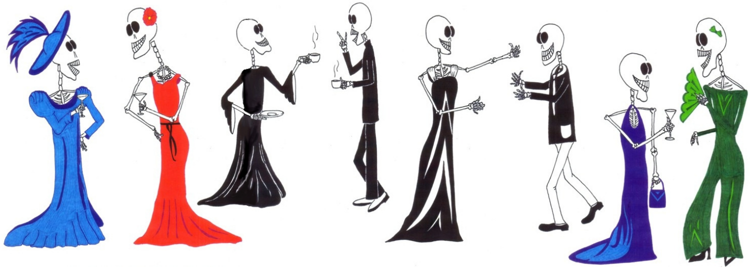
Objective:

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

In the example: I=KCCA.

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

ISA solution

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

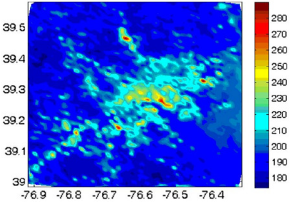
$$\mathbf{y} = \mathbf{W}\mathbf{x} = \left[\mathbf{y}^1; \ldots; \mathbf{y}^M\right],$$

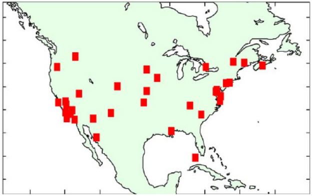
$$J(\mathbf{W}) = \mathrm{I}\left(\mathbf{y}^1, \ldots, \mathbf{y}^M\right) \to \min_{\mathbf{W}}.$$

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

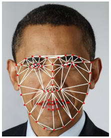
- **Goal**: aerosol prediction = air pollution → climate.

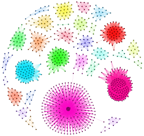
- Prediction using labelled bags:
  - bag := multi-spectral satellite measurements over an area,
  - label := local aerosol value.





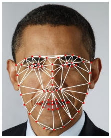
Objects in the bags

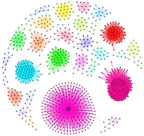




- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, . . .

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

Regression on labelled bags

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

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

Regression on labelled bags

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

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

- Prediction:

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

Regression on labelled bags

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

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

- Prediction:

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

Inner product of distributions

$K\left(\mu_{\hat{\mathbb{P}}_i}, \mu_{\hat{\mathbb{P}}_j}\right) = ?$

Feature selection

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

Feature selection: equations

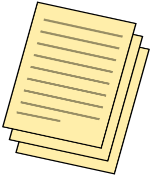
- Features: $x^1, \ldots, x^F$. Subset: $S \subseteq \{1, \ldots, F\}$.
- MaxRelevance - MinRedundancy principle [Peng et al., 2005]:

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

Hypothesis Testing

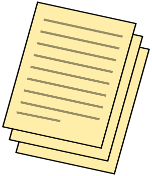
Example-1 (2-sample testing): NLP

- Given: 2 categories of documents. Examples:
  1. Bayesian inference, neuroscience.
  2. adult attachment classes.
- Task:
  - test their distinguishability,
  - most discriminative words → interpretability.



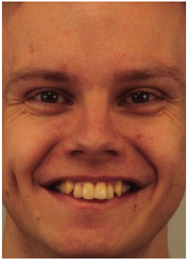
Example-1 (2-sample testing): NLP

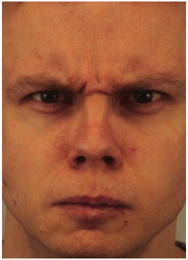
- Given: 2 categories of documents. Examples:
  1. Bayesian inference, neuroscience.
  2. adult attachment classes.
- Task:
  - test their distinguishability,
  - most discriminative words → interpretability.



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

Example-2 (2-sample testing): computer vision

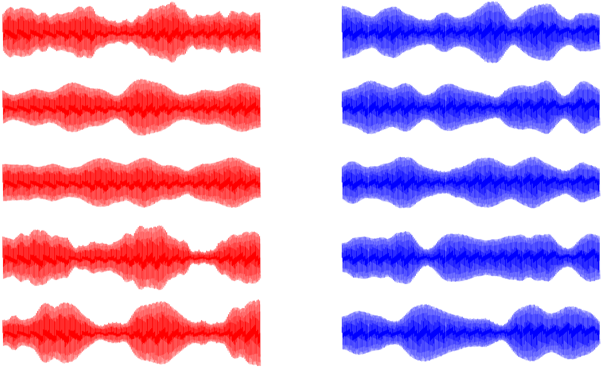




- Given: two sets of faces (happy, angry).
- Task:
  - check if they are different,
  - determine the most discriminative features/regions.

Example-3 (2-sample testing): audio

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



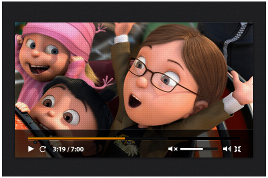
Example: independence testing

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

Example: independence testing

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

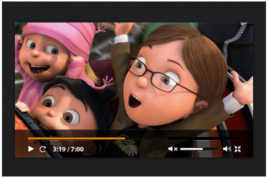
  - (video, caption) pairs



Example: independence testing

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

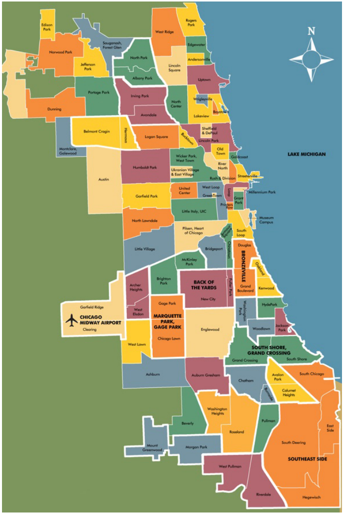
  - (video, caption) pairs



- $\{(x_i, y_i)\}_{i=1}^n \xrightarrow{?} \mathbb{P}_{xy} = \mathbb{P}_x \mathbb{P}_y$.

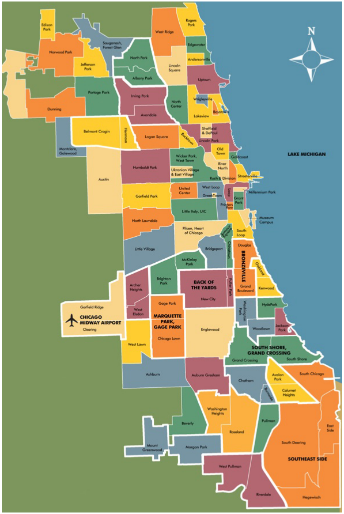
Example: goodness-of-fit testing

- Demo: criminal data analysis.
- Given:
  - Density/model: p.



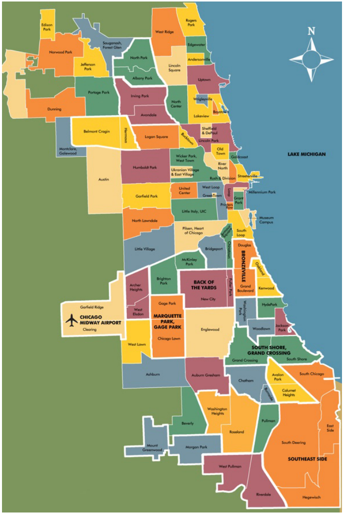
Example: goodness-of-fit testing

- Demo: criminal data analysis.
- Given:
  - Density/model: p.
  - Samples: $X = \{x_i\}_{i=1}^n \sim q$ (unknown).



- Demo: criminal data analysis.
- Given:
  - Density/model: p.
  - Samples: $X = \{x_i\}_{i=1}^n \sim q$ (unknown).
- Task: using p, X test

$$H_0 : p = q, \text{ vs}$$
$$H_1 : p \neq q.$$



'Classical' information theory

- Kullback-Leibler divergence:

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log\left[\frac{p(x)}{q(x)}\right] \mathrm{d}x.$$

'Classical' information theory

- Kullback-Leibler divergence:

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log\left[\frac{p(x)}{q(x)}\right] \mathrm{d}x.$$

- Mutual information:

$$\mathrm{I}(\mathbb{P}) = \mathrm{KL}\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right).$$

'Classical' information theory

- Kullback-Leibler divergence:

$$\mathrm{KL}\left(\mathbb{P},\mathbb{Q}\right)=\int_{\mathbb{R}^d} p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

- Mutual information:

$$\mathrm{I}\left(\mathbb{P}\right)=\mathrm{KL}\left(\mathbb{P},\otimes_{m=1}^{M}\mathbb{P}_m\right).$$

Properties:

1. $\mathrm{I}(\mathbb{P})\geqslant 0$.

'Classical' information theory

- Kullback-Leibler divergence:

$$\mathsf{KL}\left(\mathbb{P},\mathbb{Q}\right) = \int_{\mathbb{R}^d} p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

- Mutual information:

$$\mathsf{I}\left(\mathbb{P}\right) = \mathsf{KL}\left(\mathbb{P},\otimes_{m=1}^{M}\mathbb{P}_m\right).$$

Properties:

1. $\mathsf{I}(\mathbb{P}) \geqslant 0$.
2. $\mathsf{I}(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^{M}\mathbb{P}_m$.

- Kullback-Leibler divergence:

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log \left[\frac{p(x)}{q(x)} \right] \mathrm{d}x.$$

- Mutual information:

$$\mathrm{I}(\mathbb{P}) = \mathrm{KL}\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right).$$

Properties:

1. $\mathrm{I}(\mathbb{P}) \geqslant 0$.
2. $\mathrm{I}(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^{M} \mathbb{P}_m$.

Alternatives: Rényi, Tsallis, L^2 divergence... $\mathcal{X} = \mathbb{R}^d$.

Euclidean space → inner product → kernel

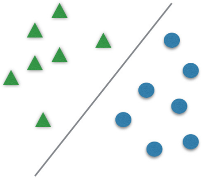
Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ leads to kernels.

Euclidean space $\rightarrow$ inner product $\rightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ leads to kernels. Why?

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ leads to kernels. Why?

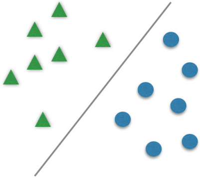
- Classification (SVM):

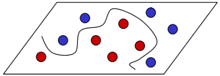


Euclidean space → inner product → kernel

Extension of $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$ leads to kernels. **Why?**

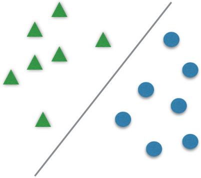
- Classification (SVM):

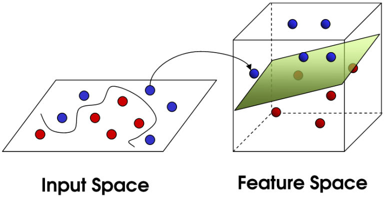




Extension of $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$ leads to kernels. Why?

- Classification (SVM):

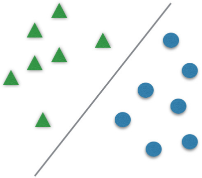


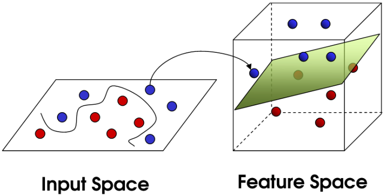


Euclidean space → inner product → kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ leads to kernels. Why?

- Classification (SVM):





- Representation of distributions:

$$\mathbb{P} \mapsto \mathbb{E}_{\mathbf{x} \sim \mathbb{P}} \varphi(\mathbf{x}).$$

Example: $\varphi(\mathbf{x}) = \mathbf{x}$: mean.

Distribution representation via functions

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

Distribution representation via functions

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

- Cdf:

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

Distribution representation via functions

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

- Cdf:

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

- Characteristic function:

$$\mathbb{P} \mapsto \mathsf{c}_{\mathbb{P}}(z) = \int_{\mathbb{R}^d} e^{i\langle z, x\rangle} \mathrm{d}\mathbb{P}(x).$$

Distribution representation via functions

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

- Cdf:

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

- Characteristic function:

$$\mathbb{P} \mapsto \mathsf{c}_{\mathbb{P}}(z) = \int_{\mathbb{R}^d} e^{i\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto \mathsf{M}_{\mathbb{P}}(z) = \int_{\mathbb{R}^d} e^{\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

Distribution representation via functions

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

- Cdf:

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

- Characteristic function:

$$\mathbb{P} \mapsto \mathrm{c}_{\mathbb{P}}(z) = \int_{\mathbb{R}^d} e^{i\langle z, x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto \mathrm{M}_{\mathbb{P}}(z) = \int_{\mathbb{R}^d} e^{\langle z, x\rangle} \mathrm{d}\mathbb{P}(x).$$

Trick

φ: on any kernel-endowed domain!

Some kernel-enriched domains: $\varphi \to k$

- 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] $\xrightarrow{\text{spec.}}$ permutations [Jiao and Vert, 2016],
- graphs [Vishwanathan et al., 2010, Kondor and Pan, 2016].

Objects of Interest

'KL divergence & mutual information' on kernel-endowed domains.

- Mean embedding:

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

Objects of Interest

'KL divergence & mutual information' on kernel-endowed domains.

- Mean embedding:

$$\mu_{\mathbb{P}} := \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \,\mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

Objects of Interest

'KL divergence & mutual information' on kernel-endowed domains.

- Mean embedding:

$$\mu_{\mathbb{P}} := \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \,\mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$\text{MMD}_k(\mathbb{P},\mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion, $k = \otimes_{m=1}^{M} k_m$:

$$\text{HSIC}_k\left(\mathbb{P}\right) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M}\mathbb{P}_m\right).$$

$\mu_{\mathbb{P}}$, MMD: review [Muandet et al., 2017]

Applications:

- two-sample testing [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Borgwardt et al., 2006, Harchaoui et al., 2007, Gretton et al., 2012, Jitkrittum et al., 2016], and its differential private variant [Raj et al., 2019]; independence [Gretton et al., 2008, Pfister et al., 2017, Jitkrittum et al., 2017a] and goodness-of-fit testing [Jitkrittum et al., 2017b, Balasubramanian et al., 2017], causal discovery [Mooij et al., 2016, Pfister et al., 2017],
- domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017], change-point detection [Harchaoui and Cappé, 2007], post selection inference [Yamada et al., 2018],
- kernel Bayesian inference [Song et al., 2011, Fukumizu et al., 2013], approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015], model criticism [Lloyd et al., 2014, Kim et al., 2016],
- topological data analysis [Kusano et al., 2016],
- distribution classification [Muandet et al., 2011, Lopez-Paz et al., 2015, Zaheer et al., 2017], distribution regression [Szabó et al., 2016, Law et al., 2018],
- generative adversarial networks [Dziugaite et al., 2015, Li et al., 2015, Binkowski et al., 2018], understanding the dynamics of complex dynamical systems [Klus et al., 2018, Klus et al., 2019], ...

MMD $\xrightarrow{\text{spec.}}$ HSIC

MMD with $k = \otimes_{m=1}^{M} k_m$:

$$\text{HSIC}_k(\mathbb{P}) = \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right).$$

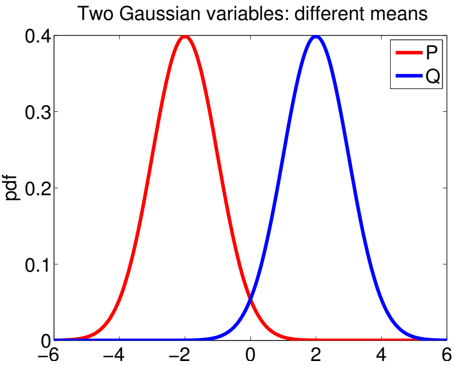
MMD with $k = \otimes_{m=1}^{M} k_m$:

$$\text{HSIC}_k(\mathbb{P}) = \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right).$$

MMD, HSIC: Easy to Estimate!

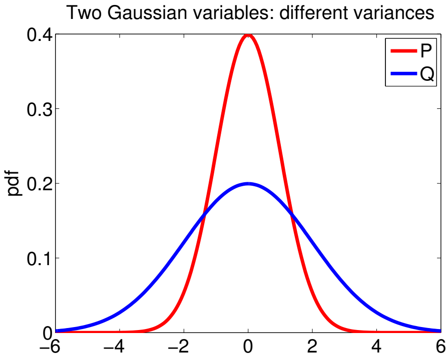
Representations of distributions: $\mathbb{E}X$

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



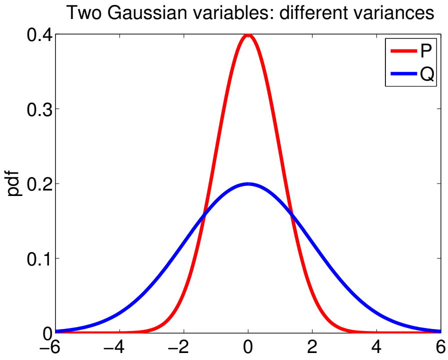
Representations of distributions: $\mathbb{E}X^2$

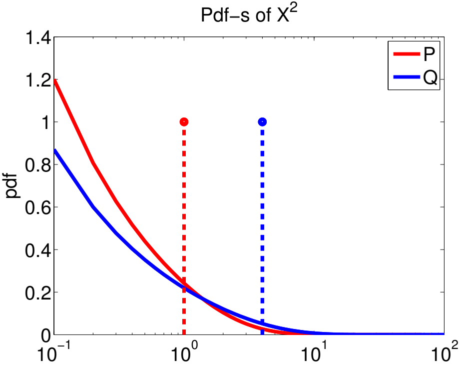
- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.



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

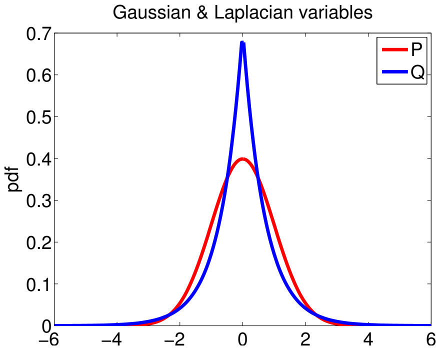
- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.
- $\varphi(x) = x^2 \Rightarrow$ difference in $\mathbb{E}X^2$.





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.



$\varphi(\mathbf{x}) = e^{i\langle \cdot, \mathbf{x} \rangle}$: characteristic function, $\mathcal{X} = \mathbb{R}^d$.

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

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

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

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

$$S = \|C_{xy}\|_F$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

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

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

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

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0 \leftrightarrow \text{ linear dependence.}$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

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

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0 \leftrightarrow \text{ linear dependence.}$$

- Covariance operator: take features of x and y

$$C_{xy} = \mathbb{E}_{xy}\Big[\underbrace{(\varphi(x) - \mathbb{E}_x\varphi(x))}_{\text{centering in feature space}} \otimes (\psi(y) - \mathbb{E}_y\psi(y)) \Big]$$

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

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

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0 \leftrightarrow \text{ linear dependence.}$$

- Covariance operator: take features of x and y

$$C_{xy} = \mathbb{E}_{xy}\Big[\underbrace{(\varphi(x) - \mathbb{E}_x\varphi(x))}_{\text{centering in feature space}} \otimes (\psi(y) - \mathbb{E}_y\psi(y))\Big],$$

$$S = \|C_{xy}\|_{HS} =: \text{HSIC}(\mathbb{P}_{xy}).$$

We capture **non-linear dependencies** via φ, ψ!

Notions we will learn about

- Kernel (k), RKHS ($\mathcal{H}_k$) $\rightarrow$ classification, regression (ridge), PCA.
- Mean embedding ($\mu_{\mathbb{P}}$): characteristic property, universality,
- $\otimes_m k_m$, $\otimes_m \mathcal{H}_{k_m}$, covariance operator,
- MMD, HSIC, KCCA,
- with applications.

Kernels & Friends

Kernel: similarity between features

- Given: x and x' objects (images or texts).

Kernel: similarity between features

- Given: x and x' objects (images or texts).
- Question: how similar they are?

Kernel: similarity between features

- Given: x and x' objects (images or texts).
- Question: how similar they are?
- Define features of the objects:

$$\varphi(x) : \text{features of } x,$$
$$\varphi(x') : \text{features of } x'.$$

- Kernel: inner product of these features

$$k(x, x') := \langle \varphi(x), \varphi(x') \rangle .$$

Kernel examples on $\mathbb{R}^d$ ($\gamma > 0, p \in \mathbb{Z}^+$)

- Polynomial kernel:

$$k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p.$$

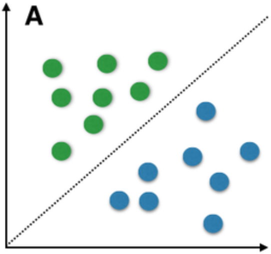
- Gaussian kernel:

$$k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}.$$

Non-linear features: why?

Classification motivation: linear separability

Idealized situation

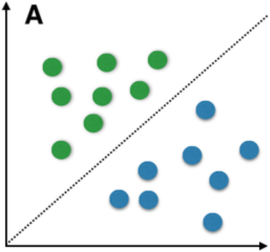


Decision surface:

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle = 0\}$$

Classification motivation: linear separability

Idealized situation



Decision surface:

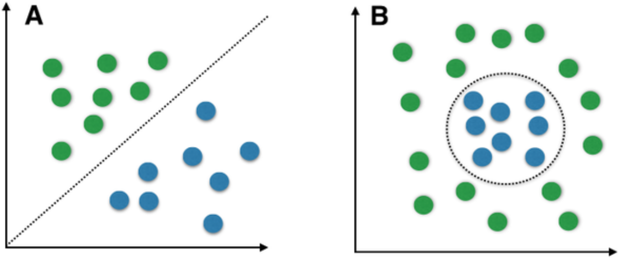
$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle = 0\} \Rightarrow$$

classes:

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle \geqslant 0\} \qquad \{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle < 0\}$$

Classification motivation: non-linear separability

Idealized situation | Real world



Decision surface (left):

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle = 0\} \Rightarrow$$

classes:

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle \geqslant 0\} \qquad \{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle < 0\}.$$

Non-linear separability – continued

On the ellipse

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\}$$

Non-linear separability – continued

On the ellipse, outside

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\},$$
$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1\right\}$$

Non-linear separability – continued

On the ellipse, outside, inside:

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\},$$
$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1\right\},$$
$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} < 1\right\}.$$

Non-linear separability – continued

On the ellipse, outside, inside:

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} < 1\right\}.$$

With polynomial feature: $\varphi(\mathbf{x}) = (x_1^2, x_1, 1, x_2^2, x_2)$:

- Decision surface: $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x})\rangle = 0\}$.

Non-linear separability – continued

On the ellipse, outside, inside:

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} < 1\right\}.$$

With polynomial feature: $\varphi(\mathbf{x}) = (x_1^2, x_1, 1, x_2^2, x_2)$:

- Decision surface: $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle = 0\}$.
- Classes: $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle > 0\}$, $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle < 0\}$.

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$ $=?$

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\begin{aligned}
\varphi(\mathbf{x}) &= \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),\\
\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle \\
&= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2}x_1x_2(x_1')(x_2') + x_2^2(x_2')^2 \\
&= \left(x_1x_1' + x_2x_2'\right)^2
\end{aligned}$$

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2}\, x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

$$= \left(x_1x_1' + x_2x_2'\right)^2$$

$$= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \langle \mathbf{x}, \mathbf{x}' \rangle^2 =: k(\mathbf{x}, \mathbf{x}').$$

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2}\, x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

$$= \left(x_1x_1' + x_2x_2'\right)^2$$

$$= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \langle\mathbf{x}, \mathbf{x}'\rangle^2 =: k(\mathbf{x}, \mathbf{x}').$$

$\langle\mathbf{x}, \mathbf{x}'\rangle^d = \langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle$: $\varphi(\mathbf{x}) =$ d-order polynomial. $\Rightarrow$

Quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

$$= \left(x_1x_1' + x_2x_2'\right)^2$$

$$= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \langle \mathbf{x}, \mathbf{x}' \rangle^2 =: k(\mathbf{x}, \mathbf{x}').$$

$\langle \mathbf{x}, \mathbf{x}' \rangle^d = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$: $\varphi(\mathbf{x}) =$ d-order polynomial. $\Rightarrow$ Explicit computation would be heavy!

Maximum correlation: KCCA

- 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.
- Idea:

$$Q(\mathbb{P}_{xy}) = \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \operatorname{corr}(f(x), g(y))$$

Maximum correlation: KCCA

- 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.
- Idea:

$$Q(\mathbb{P}_{xy}) = \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \operatorname{corr}(f(x), g(y)),$$

- $\mathcal{F} = C_b(\mathcal{X})$, $\mathcal{G} = C_b(\mathcal{Y})$: enough, but computationally...

Maximum correlation: KCCA

- 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.
- Idea:

$$Q(\mathbb{P}_{xy}) = \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \mathrm{corr}(f(x), g(y)),$$

- $\mathcal{F} = C_b(\mathcal{X})$, $\mathcal{G} = C_b(\mathcal{Y})$: enough, but computationally...
- Trick: $\mathcal{H}_k$ dense in $C_b(\mathcal{X})$, similarly $\mathcal{H}_\ell$ dense in $C_b(\mathcal{Y})$.
  - This universality: captures independence.
  - Computationally tractable.

Kernels: why?

Linear $\to$ non-linear transition:

- Inner product of non-linear features: $k(x, x')$, implicit (highD: ✓).

Kernels: why?

Linear $\rightarrow$ non-linear transition:

- Inner product of non-linear features: $k(x, x')$, implicit (highD: ✓).
- Classification: non-linear separability.

Kernels: why?

Linear $\rightarrow$ non-linear transition:

- Inner product of non-linear features: $k(x, x')$, implicit (highD: ✓).
- Classification: non-linear separability.
- Information theory:
  - $\sup_{f,g} \operatorname{corr}(f(x), g(y))$: can capture independence (KCCA),
  - $\mathbb{E}$['rich' features] $\leftarrow$ encodes distributions (MMD).

Kernels: why?

Linear $\rightarrow$ non-linear transition:

- Inner product of non-linear features: $k(x, x')$, implicit (highD: ✓).
- Classification: non-linear separability.
- Information theory:
  - $\sup_{f,g} \operatorname{corr}(f(x), g(y))$: can capture independence (KCCA),
  - $\mathbb{E}$['rich' features] $\leftarrow$ encodes distributions (MMD).
- Computational tractability & accelerations.

Kernels: why?

Linear $\rightarrow$ non-linear transition:

- Inner product of non-linear features: $k(x, x')$, implicit (highD: ✓).
- Classification: non-linear separability.
- Information theory:
  - $\sup_{f,g} \mathrm{corr}(f(x), g(y))$: can capture independence (KCCA),
  - $\mathbb{E}$['rich' features] $\leftarrow$ encodes distributions (MMD).
- Computational tractability & accelerations.
- Hilbert space: enables analysis.

Code

- ITE toolbox (KL, MI, MMD, HSIC, KCCA & many more):

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

Code

- ITE toolbox (KL, MI, MMD, HSIC, KCCA & many more):

  `https://bitbucket.org/szzoli/ite-in-python/`
- Linear-time hypothesis testing
  - two-sample (NIPS-2016, oral):

    `https://github.com/wittawatj/interpretable-test`

Code

- ITE toolbox (KL, MI, MMD, HSIC, KCCA & many more):

  `https://bitbucket.org/szzoli/ite-in-python/`
- Linear-time hypothesis testing
  - two-sample (NIPS-2016, oral):

    `https://github.com/wittawatj/interpretable-test`
  - independence (ICML-2017):

    `.../fsic-test`

Code

- ITE toolbox (KL, MI, MMD, HSIC, KCCA & many more):

  `https://bitbucket.org/szzoli/ite-in-python/`
- Linear-time hypothesis testing
  - two-sample (NIPS-2016, oral):

    `https://github.com/wittawatj/interpretable-test`
  - independence (ICML-2017):

    `.../fsic-test`
  - goodness-of-fit (NIPS-2017, best paper award):

    `.../kernel-gof`

Kernels

Kernel, RKHS

- Given: $\mathcal{X}$ set.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called kernel if $\exists \varphi : \mathcal{X} \to \mathcal{H}$ilbert space

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

Kernel, RKHS

- Given: $\mathcal{X}$ set.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called kernel if $\exists \varphi : \mathcal{X} \to \mathcal{H}$ilbert space

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- $\varphi / \mathcal{H}$ might not be unique: $k(x, y) = xy$

$$\varphi_1(x) = x \in \mathbb{R}, \qquad \varphi_2(x) = \frac{1}{\sqrt{2}}(x, x) \in \mathbb{R}^2.$$

Kernel, RKHS

- Given: $\mathcal{X}$ set.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called kernel if $\exists \varphi : \mathcal{X} \to \mathcal{H}$ilbert space

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- $\varphi/\mathcal{H}$ might not be unique: $k(x, y) = xy$

$$\varphi_1(x) = x \in \mathbb{R}, \qquad \varphi_2(x) = \frac{1}{\sqrt{2}}(x, x) \in \mathbb{R}^2.$$

- $\mathcal{H}$ intuition: vectors, inner product, complete ('no holes').

A bit of functional analysis follows $\approx$ linalg, geometry!

Vector space: $(V, +, \lambda\cdot)$

- Points $=$ vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), \quad (\text{associativity})$$

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), \quad \text{(associativity)}$$
$$\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1, \quad \text{(commutativity)}$$

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), \quad \text{(associativity)}$$
$$\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1, \quad \text{(commutativity)}$$
$$\exists \mathbf{0} : \mathbf{v} + \mathbf{0} = \mathbf{v},$$

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$
\begin{aligned}
(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 &= \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), && \text{(associativity)}\\
\mathbf{v}_1 + \mathbf{v}_2 &= \mathbf{v}_2 + \mathbf{v}_1, && \text{(commutativity)}\\
\exists \mathbf{0} : \mathbf{v} + \mathbf{0} &= \mathbf{v},\\
\exists -\mathbf{v} : \mathbf{v} + (-\mathbf{v}) &= \mathbf{0},
\end{aligned}
$$

Vector space: $(V, +, \lambda\cdot)$

- Points $=$ vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda \mathbf{v}$ with the 'natural' properties

$$
\begin{aligned}
(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 &= \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), && \text{(associativity)}\\
\mathbf{v}_1 + \mathbf{v}_2 &= \mathbf{v}_2 + \mathbf{v}_1, && \text{(commutativity)}\\
\exists \mathbf{0} : \mathbf{v} + \mathbf{0} &= \mathbf{v}, \\
\exists - \mathbf{v} : \mathbf{v} + (-\mathbf{v}) &= \mathbf{0}, \\
a(b\mathbf{v}) &= (ab)\mathbf{v},
\end{aligned}
$$

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$
\begin{aligned}
(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 &= \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), && \text{(associativity)} \\
\mathbf{v}_1 + \mathbf{v}_2 &= \mathbf{v}_2 + \mathbf{v}_1, && \text{(commutativity)} \\
\exists \mathbf{0} : \mathbf{v} + \mathbf{0} &= \mathbf{v}, \\
\exists -\mathbf{v} : \mathbf{v} + (-\mathbf{v}) &= \mathbf{0}, \\
a(b\mathbf{v}) &= (ab)\mathbf{v}, \\
1\mathbf{v} &= \mathbf{v},
\end{aligned}
$$

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$
\begin{aligned}
(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 &= \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), && \text{(associativity)}\\
\mathbf{v}_1 + \mathbf{v}_2 &= \mathbf{v}_2 + \mathbf{v}_1, && \text{(commutativity)}\\
\exists \mathbf{0} : \mathbf{v} + \mathbf{0} &= \mathbf{v}, \\
\exists -\mathbf{v} : \mathbf{v} + (-\mathbf{v}) &= \mathbf{0}, \\
a(b\mathbf{v}) &= (ab)\mathbf{v}, \\
1\mathbf{v} &= \mathbf{v}, \\
a(\mathbf{v}_1 + \mathbf{v}_2) &= a\mathbf{v}_1 + a\mathbf{v}_2,
\end{aligned}
$$

Vector space: $(V, +, \lambda\cdot)$

- Points = vectors.
- 2 operations: $\mathbf{v}_1 + \mathbf{v}_2$, $\lambda\mathbf{v}$ with the 'natural' properties

$$
\begin{aligned}
(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 &= \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3), \quad \text{(associativity)}\\
\mathbf{v}_1 + \mathbf{v}_2 &= \mathbf{v}_2 + \mathbf{v}_1, \quad \text{(commutativity)}\\
\exists \mathbf{0} : \mathbf{v} + \mathbf{0} &= \mathbf{v},\\
\exists -\mathbf{v} : \mathbf{v} + (-\mathbf{v}) &= \mathbf{0},\\
a(b\mathbf{v}) &= (ab)\mathbf{v},\\
1\mathbf{v} &= \mathbf{v},\\
a(\mathbf{v}_1 + \mathbf{v}_2) &= a\mathbf{v}_1 + a\mathbf{v}_2,\\
(a + b)\mathbf{v} &= a\mathbf{v} + b\mathbf{v} \qquad \text{for } \forall \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v} \in V, a, b \in \mathbb{R}.
\end{aligned}
$$

Vector space: examples

1. $(\mathbb{R}^d, +, \cdot)$ defined as

$$(x_1, \ldots, x_d) + (y_1, \ldots, y_d) := (x_1 + y_1, \ldots, x_d + y_d),$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

Vector space: examples

1. $(\mathbb{R}^d, +, \cdot)$ defined as

$$(x_1, \ldots, x_d) + (y_1, \ldots, y_d) := (x_1 + y_1, \ldots, x_d + y_d),$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

2. Polynomials:

$$\sum_i a_i x^i + \sum_i b_i x^i := \sum_i (a_i + b_i) x^i,$$
$$\lambda \cdot \left(\sum_i a_i x^i \right) := \sum_i (\lambda a_i) x^i.$$

Vector space: examples

1. $(\mathbb{R}^d, +, \cdot)$ defined as

$$(x_1, \ldots, x_d) + (y_1, \ldots, y_d) := (x_1 + y_1, \ldots, x_d + y_d),$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

2. Polynomials:

$$\sum_i a_i x^i + \sum_i b_i x^i := \sum_i (a_i + b_i) x^i,$$
$$\lambda \cdot \left(\sum_i a_i x^i \right) := \sum_i (\lambda a_i) x^i.$$

3. $\mathcal{X} \to \mathbb{R}$-valued functions:

$$(f + g)(x) := f(x) + g(x), \qquad (\lambda \cdot f)(x) := \lambda f(x).$$

Vector space: examples

1. $(\mathbb{R}^d, +, \cdot)$ defined as

$$(x_1, \dots, x_d) + (y_1, \dots, y_d) := (x_1 + y_1, \dots, x_d + y_d),$$
$$\lambda \cdot (x_1, \dots, x_d) := (\lambda x_1, \dots, \lambda x_d).$$

2. Polynomials:

$$\sum_i a_i x^i + \sum_i b_i x^i := \sum_i (a_i + b_i) x^i,$$
$$\lambda \cdot \left(\sum_i a_i x^i \right) := \sum_i (\lambda a_i) x^i.$$

3. $\mathcal{X} \to \mathbb{R}$-valued functions:

$$(f + g)(x) := f(x) + g(x), \qquad (\lambda \cdot f)(x) := \lambda f(x).$$

Previously: $\mathcal{X} = \{1, \dots, d\}$, $\mathcal{X} = \mathbb{N}$.

Vector space: examples – continued

4. $\supset C[a, b]$: continuous functions.

Vector space: examples – continued

4. $\supset C[a, b]$: continuous functions.
5. $(\mathbb{R}^d, \otimes, \cdot)$:

$$(x_1, \ldots, x_d) \otimes (y_1, \ldots, y_d) := [x_i y_j]_{i,j=1}^d$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

Vector space: examples – continued

4. $\supset C[a, b]$: continuous functions.
5. $(\mathbb{R}^d, \otimes, \cdot)$: NO!!!

$$(x_1, \ldots, x_d) \otimes (y_1, \ldots, y_d) := [x_i y_j]_{i,j=1}^d \in \mathbb{R}^{d \times d} \neq \mathbb{R}^d,$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

Vector space: examples – continued

4. $\supset C[a,b]$: continuous functions.
5. $(\mathbb{R}^d, \otimes, \cdot)$: NO!!!

$$(x_1,\ldots,x_d) \otimes (y_1,\ldots,y_d) := [x_i y_j]_{i,j=1}^d \in \mathbb{R}^{d\times d} \neq \mathbb{R}^d,$$
$$\lambda \cdot (x_1,\ldots,x_d) := (\lambda x_1,\ldots,\lambda x_d).$$

6. $(\mathbb{R}^d, \langle\cdot,\cdot\rangle, \cdot)$: similarly NO

$$\langle (x_1,\ldots,x_d), (y_1,\ldots,y_d)\rangle := \sum_{i=1}^d x_i y_i \in \mathbb{R} \neq \mathbb{R}^d,$$
$$\lambda \cdot (x_1,\ldots,x_d) := (\lambda x_1,\ldots,\lambda x_d).$$

Vector space: examples – continued

4. $\supset C[a, b]$: continuous functions.
5. $(\mathbb{R}^d, \otimes, \cdot)$: NO!!!

$$(x_1, \ldots, x_d) \otimes (y_1, \ldots, y_d) := [x_i y_j]_{i,j=1}^d \in \mathbb{R}^{d \times d} \neq \mathbb{R}^d,$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

6. $(\mathbb{R}^d, \langle \cdot, \cdot \rangle, \cdot)$: similarly NO

$$\langle (x_1, \ldots, x_d), (y_1, \ldots, y_d) \rangle := \sum_{i=1}^d x_i y_i \in \mathbb{R} \neq \mathbb{R}^d,$$
$$\lambda \cdot (x_1, \ldots, x_d) := (\lambda x_1, \ldots, \lambda x_d).$$

Now we put a notion of norm, inner product to vectors.

Normed space

We define the 'length' of a vector.

$\mathcal{H}$: vector space over $\mathbb{R}$. $\|\cdot\| : \mathcal{H} \to [0, \infty)$ is **norm** on $\mathcal{H}$, if

1. $\|f\| = 0$ iff. $f = 0$ (norm separates points),

Normed space

We define the '**length**' of a vector.

$\mathcal{H}$: vector space over $\mathbb{R}$. $\|\cdot\| : \mathcal{H} \to [0, \infty)$ is **norm** on $\mathcal{H}$, if

1. $\|f\| = 0$ iff. $f = 0$ (norm separates points),
2. $\|\lambda f\| = |\lambda| \|f\|$ $\forall \lambda \in \mathbb{R}, \forall f \in \mathcal{H}$ (positive homogenity),

Normed space

We define the 'length' of a vector.

$\mathcal{H}$: vector space over $\mathbb{R}$. $\|\cdot\| : \mathcal{H} \to [0, \infty)$ is norm on $\mathcal{H}$, if

1. $\|f\| = 0$ iff. $f = 0$ (norm separates points),
2. $\|\lambda f\| = |\lambda| \, \|f\| \; \forall \lambda \in \mathbb{R}, \forall f \in \mathcal{H}$ (positive homogenity),
3. $\|f + g\| \leqslant \|f\| + \|g\| \; \forall f, g \in \mathcal{H}$ (triangle inequality).

Normed space

We define the 'length' of a vector.

$\mathcal{H}$: vector space over $\mathbb{R}$. $\|\cdot\| : \mathcal{H} \to [0, \infty)$ is **norm** on $\mathcal{H}$, if

1. $\|f\| = 0$ iff. $f = 0$ (norm separates points),
2. $\|\lambda f\| = |\lambda| \|f\| \; \forall \lambda \in \mathbb{R}, \forall f \in \mathcal{H}$ (positive homogenity),
3. $\|f + g\| \leq \|f\| + \|g\| \; \forall f, g \in \mathcal{H}$ (triangle inequality).

Note:

- norm $\Rightarrow$ metric: $\rho(f, g) = \|f - g\| \Rightarrow$
- study continuity, convergence.

Normed space: examples

- $(\mathbb{R}, |\cdot|)$.

Normed space: examples

- $(\mathbb{R}, |\cdot|)$.
- $\left(\mathbb{R}^d, \|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}\right)$, $p \in [0, \infty]$.
  - $p = 2$: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ (Euclidean),

Normed space: examples

- $(\mathbb{R}, |\cdot|)$.
- $\left(\mathbb{R}^d, \|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}\right)$, $p \in [0, \infty]$.
  - $p = 2$: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ (Euclidean),
  - $p = 1$: $\|\mathbf{x}\|_1 = \sum_i |x_i|$ (Manhattan),

Normed space: examples

- $(\mathbb{R}, |\cdot|)$.
- $\left(\mathbb{R}^d, \|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}\right)$, $p \in [0, \infty]$.
  - $p = 2$: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ (Euclidean),
  - $p = 1$: $\|\mathbf{x}\|_1 = \sum_i |x_i|$ (Manhattan),
  - $p = \infty$: $\|\mathbf{x}\|_\infty = \max_i |x_i|$ (maximum norm).

Normed space: examples

- $(\mathbb{R}, |\cdot|)$.
- $\left(\mathbb{R}^d, \|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}\right)$, $p \in [0, \infty]$.
  - $p = 2$: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ (Euclidean),
  - $p = 1$: $\|\mathbf{x}\|_1 = \sum_i |x_i|$ (Manhattan),
  - $p = \infty$: $\|\mathbf{x}\|_\infty = \max_i |x_i|$ (maximum norm).
- $\left(C[a, b], \|f\|_p = \left[\int_a^b |f(x)|^p \mathrm{d}x\right]^{\frac{1}{p}}\right)$, $1 \leqslant p < \infty$.

Normed space: examples

- $(\mathbb{R}, |\cdot|)$.
- $\left(\mathbb{R}^d, \|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}\right)$, $p \in [0, \infty]$.
  - $p = 2$: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ (Euclidean),
  - $p = 1$: $\|\mathbf{x}\|_1 = \sum_i |x_i|$ (Manhattan),
  - $p = \infty$: $\|\mathbf{x}\|_\infty = \max_i |x_i|$ (maximum norm).
- $\left(C[a,b], \|f\|_p = \left[\int_a^b |f(x)|^p \mathrm{d}x\right]^{\frac{1}{p}}\right)$, $1 \leqslant p < \infty$.
- $\left(C[a,b], \|f\|_\infty = \max_{x\in[a,b]} |f(x)|\right)$, $p = \infty$.

Inner product space (also called Euclidean space)

$\mathcal{H}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{H}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),

Inner product space (also called Euclidean space)

$\mathcal{H}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an **inner product** on $\mathcal{H}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{H}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),

Inner product space (also called Euclidean space)

$\mathcal{H}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{H}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),
3. $\langle f, f \rangle \geqslant 0$; $\langle f, f \rangle = 0 \Leftrightarrow f = 0$.

Inner product space (also called Euclidean space)

$\mathcal{H}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an **inner product** on $\mathcal{H}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{H}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),
3. $\langle f, f \rangle \geqslant 0$; $\langle f, f \rangle = 0 \Leftrightarrow f = 0$.

Notes:

- 1, 2 $\Rightarrow$ bilinearity.

Inner product space (also called Euclidean space)

$\mathcal{H}$: vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an **inner product** on $\mathcal{H}$ if for $\forall \alpha_i \in \mathbb{R}, f_i, f, g \in \mathcal{H}$

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$ (linearity),
2. $\langle f, g \rangle = \langle g, f \rangle$ (symmetry),
3. $\langle f, f \rangle \geqslant 0$; $\langle f, f \rangle = 0 \Leftrightarrow f = 0$.

Notes:

- 1, 2 $\Rightarrow$ bilinearity. Inner product $\Rightarrow$

$$\text{norm: } \|f\| = \sqrt{\langle f, f \rangle}, \qquad \text{angle: } \cos(f, g) = \frac{\langle f, g \rangle}{\|f\| \, \|g\|}.$$

Inner product space: examples

- $\left(\mathbb{R}^d, \langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^d x_i y_i\right)$.
- $\left(\mathbb{R}^{d_1 \times d_2}, \langle \mathbf{A}, \mathbf{B} \rangle_F = \text{tr}(\mathbf{A}^T \mathbf{B}) = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} A_{ij} B_{ij}\right)$.
- $\left(C[a,b], \langle f, g \rangle = \int_a^b f(x) g(x) \mathrm{d}x\right)$.

Norm vs inner product

Relations:

- $|\langle f, g\rangle| \leqslant \|f\| \cdot \|g\|$ (CBS),

Norm vs inner product

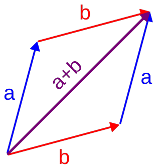
Relations:

- $|\langle f, g\rangle| \leqslant \|f\| \cdot \|g\|$ (CBS),
- $4\langle f, g\rangle = \|f+g\|^2 - \|f-g\|^2$ (polarization identity),

Norm vs inner product

Relations:

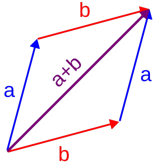
- $|\langle f, g\rangle| \leqslant \|f\| \cdot \|g\|$ (CBS),
- $4\langle f, g\rangle = \|f+g\|^2 - \|f-g\|^2$ (polarization identity),
- $\|a+b\|^2 + \|a-b\|^2 = (*)$ (parallelogram rule).



Norm vs inner product

Relations:

- $|\langle f, g\rangle| \leqslant \|f\| \cdot \|g\|$ (CBS),
- $4\langle f, g\rangle = \|f+g\|^2 - \|f-g\|^2$ (polarization identity),
- $\|a+b\|^2 + \|a-b\|^2 = (*)$ (parallelogram rule).

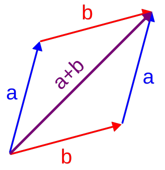


$$(*) = \langle a+b, a+b\rangle + \langle a-b, a-b\rangle$$

Norm vs inner product

Relations:

- $|\langle f, g\rangle| \leqslant \|f\| \cdot \|g\|$ (CBS),
- $4\langle f, g\rangle = \|f+g\|^2 - \|f-g\|^2$ (polarization identity),
- $\|a+b\|^2 + \|a-b\|^2 = (*)$ (parallelogram rule).

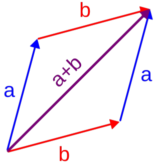


$$(*) = \langle a+b, a+b\rangle + \langle a-b, a-b\rangle$$
$$= 2\left(\|a\|^2 + \|b\|^2\right) \pm 2\langle a, b\rangle$$

Norm vs inner product

Relations:

- $|\langle f, g\rangle| \leqslant \|f\| \cdot \|g\|$ (CBS),
- $4\langle f, g\rangle = \|f+g\|^2 - \|f-g\|^2$ (polarization identity),
- $\|a+b\|^2 + \|a-b\|^2 = (*)$ (parallelogram rule).



$$(*) = \langle a+b, a+b\rangle + \langle a-b, a-b\rangle$$
$$= 2\left(\|a\|^2 + \|b\|^2\right) \pm 2\langle a, b\rangle = 2\|a\|^2 + 2\|b\|^2.$$

The parallelogram rule holds in an inner product space.

Example when the parallelogram rule fails

$C[0,1]$ with $\|f\|_\infty = \max_{x\in[0,1]} |f(x)|$:

$$f(x) := 1 - x, \; g(x) := x$$

Example when the parallelogram rule fails

$C[0,1]$ with $\|f\|_\infty = \max_{x\in[0,1]} |f(x)|$:

$$f(x) := 1 - x, \ g(x) := x,$$
$$\|f+g\|_\infty^2 + \|f-g\|_\infty^2 = \|1\|_\infty^2 + \|\underbrace{1-2x}_{\in[-1,1]}\|_\infty^2 = 1^2 + 1^2 = 2,$$

Example when the parallelogram rule fails

$C[0,1]$ with $\|f\|_\infty = \max_{x\in[0,1]} |f(x)|$:

$$f(x) := 1 - x,\ g(x) := x,$$

$$\|f+g\|_\infty^2 + \|f-g\|_\infty^2 = \|1\|_\infty^2 + \|\underbrace{1-2x}_{\in[-1,1]}\|_\infty^2 = 1^2 + 1^2 = 2,$$

$$2\left(\|f\|_\infty^2 + \|g\|_\infty^2\right) = 2\|\underbrace{1-x}_{\in[0,1]}\|_\infty^2 + 2\|\underbrace{x}_{\in[0,1]}\|_\infty^2 = 2 + 2 = 4.$$

Example when the parallelogram rule fails

$C[0,1]$ with $\|f\|_\infty = \max_{x\in[0,1]} |f(x)|$:

$$f(x) := 1 - x, \; g(x) := x,$$
$$\|f+g\|_\infty^2 + \|f-g\|_\infty^2 = \|1\|_\infty^2 + \|\underbrace{1-2x}_{\in[-1,1]}\|_\infty^2 = 1^2 + 1^2 = 2,$$
$$2\left(\|f\|_\infty^2 + \|g\|_\infty^2\right) = 2\|\underbrace{1-x}_{\in[0,1]}\|_\infty^2 + 2\|\underbrace{x}_{\in[0,1]}\|_\infty^2 = 2+2 = 4.$$

Characterization

A norm is induced by an inner product iff

$$\|f+g\|^2 + \|f-g\|^2 = 2\left(\|f\|^2 + \|g\|^2\right) \quad \forall f,g.$$

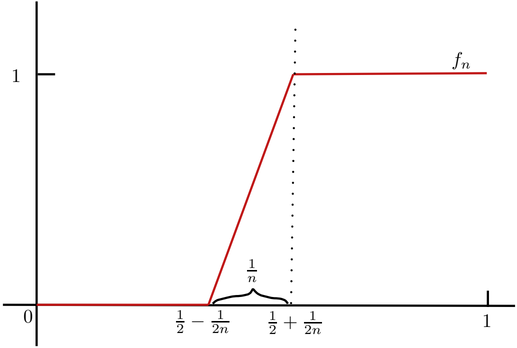
Completeness: motivation

- $\{(f_n)\}_{n\in\mathbb{N}} \subset \mathbb{Q}$: $1, 1.4, 1.41, 1.414, 1.4142, \ldots$
  - $|f_n - f_m|$ can be arbitrary small for large enough (n, m) $\leftrightarrow$ Cauchy seq,

Completeness: motivation

- $\{(f_n)\}_{n\in\mathbb{N}} \subset \mathbb{Q}$: $1, 1.4, 1.41, 1.414, 1.4142, \ldots$
  - $|f_n - f_m|$ can be arbitrary small for large enough $(n, m) \leftrightarrow$ Cauchy seq,
  - but (f_n) does not converge in $\mathbb{Q} \not\ni \sqrt{2}$.

- $\{(f_n)\}_{n\in\mathbb{N}} \subset \mathbb{Q}$: $1, 1.4, 1.41, 1.414, 1.4142, \ldots$
  - $|f_n - f_m|$ can be arbitrary small for large enough $(n, m) \leftrightarrow$ Cauchy seq,
  - but (f_n) does not converge in $\mathbb{Q} \not\ni \sqrt{2}$.
- $\left(C[0,1], \|\cdot\|_{L^2[0,1]}\right)$: similarly



Hilbert space

- $\mathcal{H}$ilbert space := complete Euclidean space. Prototype:

$$L^2(\mathcal{X},\mu) := \left\{ f : \mathcal{X} \to \mathbb{R} \ : \|f\|_2 = \left[\int_{\mathcal{X}} |f(x)|^2 \mathrm{d}\mu(x)\right]^{1/2} < \infty \right\}.$$

Hilbert space

- $\mathcal{H}$ilbert space := complete Euclidean space. Prototype:

$$L^2(\mathcal{X},\mu) := \left\{ f : \mathcal{X} \to \mathbb{R} \ : \|f\|_2 = \left[\int_{\mathcal{X}} |f(x)|^2 \mathrm{d}\mu(x)\right]^{1/2} < \infty \right\}.$$

Specifically:

$$(\mathbb{R}^d, \|\cdot\|_2), \text{ or } \ell^2(\mathbb{N}) = \left\{ (a_n)_{n\in\mathbb{N}} \ : \sqrt{\sum_{n=1}^{\infty} a_n^2} < \infty \right\}.$$

Recall: $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$.

Hilbert space

- $\mathcal{H}$ilbert space := complete Euclidean space. Prototype:

$$L^2(\mathcal{X},\mu) := \left\{ f : \mathcal{X} \to \mathbb{R} \ : \|f\|_2 = \left[\int_{\mathcal{X}} |f(x)|^2 \mathrm{d}\mu(x)\right]^{1/2} < \infty \right\}.$$

Specifically:

$$(\mathbb{R}^d, \|\cdot\|_2), \text{ or } \ell^2(\mathbb{N}) = \left\{ (a_n)_{n\in\mathbb{N}} \ : \sqrt{\sum_{n=1}^{\infty} a_n^2} < \infty \right\}.$$

Recall: $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$.

- Banach space := complete normed space:

$$L^p(\mathcal{X},\mu), \qquad (\mathbb{R}^d, \|\cdot\|_p), \qquad \ell^p(\mathbb{N}), \qquad (C[a,b], \|\cdot\|_\infty).$$

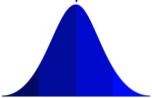
Kernels, RKHS: Definition-2

- Def-1 $=$ feature space point of view, $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}$.

Kernels, RKHS: Definition-2

- Def-1 = feature space point of view, $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}$.
- Def-2 = constructive. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of a $\mathcal{H}$(ilbert) $\subset \mathbb{R}^{\mathcal{X}}$

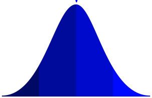
$$k(\cdot, b) \in \mathcal{H}, \qquad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



Kernels, RKHS: Definition-2

- Def-1 = feature space point of view, $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}$.
- Def-2 = constructive. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** of a $\mathcal{H}$(ilbert) $\subset \mathbb{R}^{\mathcal{X}}$

$$\underbrace{k(\cdot, b)}\in \mathcal{H}, \qquad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$

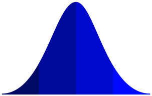


$$\xrightarrow{\text{spec.}} k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}.$$

Kernels, RKHS: Definition-2

- Def-1 = feature space point of view, $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}$.
- Def-2 = constructive. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** of a $\mathcal{H}(\text{ilbert}) \subset \mathbb{R}^{\mathcal{X}}$

$$k(\cdot, b) \in \mathcal{H}, \qquad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



$$\xrightarrow{\text{spec.}} k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}. \quad \boxed{\mathcal{H}_k = \overline{\left\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)\right\}}}.$$

Kernels: Definition-3

- Def-3: Gram matrix, optimization point of view.
- Intuition: $\mathcal{X} := \mathbb{R}^d$, data matrix $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$, then

$$\mathbf{G} := \mathbf{X}^T \mathbf{X} = \left[\langle x_i, x_j \rangle_2\right]_{i,j=1}^n \succeq 0.$$

Kernels: Definition-3

- Def-3: Gram matrix, optimization point of view.
- Intuition: $\mathcal{X} := \mathbb{R}^d$, data matrix $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$, then

$$\mathbf{G} := \mathbf{X}^T\mathbf{X} = \left[\langle x_i, x_j \rangle_2\right]_{i,j=1}^n \succeq 0.$$

i.e.

$$\mathbf{G}^T = \mathbf{G} \qquad \text{(symmetry)},$$

Kernels: Definition-3

- Def-3: Gram matrix, optimization point of view.
- Intuition: $\mathcal{X} := \mathbb{R}^d$, data matrix $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$, then

$$\mathbf{G} := \mathbf{X}^T\mathbf{X} = \left[\langle x_i, x_j \rangle_2\right]_{i,j=1}^n \succeq 0.$$

i.e.

$$\begin{aligned} \mathbf{G}^T &= \mathbf{G} && \text{(symmetry)}, \\ \mathbf{v}^T\mathbf{G}\mathbf{v} &= \mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} = \|\mathbf{X}\mathbf{v}\|_2^2 \geqslant 0 && (\forall \mathbf{v} \in \mathbb{R}^d). \end{aligned}$$

- Def-3: Gram matrix, optimization point of view.
- Intuition: $\mathcal{X} := \mathbb{R}^d$, data matrix $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$, then

$$\mathbf{G} := \mathbf{X}^T\mathbf{X} = \left[\langle x_i, x_j \rangle_2\right]_{i,j=1}^n \succeq 0.$$

i.e.

$$\mathbf{G}^T = \mathbf{G} \qquad \text{(symmetry)},$$
$$\mathbf{v}^T\mathbf{G}\mathbf{v} = \mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} = \|\mathbf{X}\mathbf{v}\|_2^2 \geqslant 0 \quad (\forall \mathbf{v} \in \mathbb{R}^d).$$

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric is positive definite if

$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \succeq 0 \quad \forall n \in \mathbb{Z}^+, \forall \{x_i\}_{i=1}^n.$$

- Def-4 intuition: We want

$$(f_n)_{n\in\mathbb{N}} \xrightarrow{\|\cdot\|} f \quad \Rightarrow \quad (f_n(x))_{n\in\mathbb{N}} \to f(x) \quad \forall x.$$

Kernels: Definition-4 – motivation

- Def-4 intuition: We want

$$(f_n)_{n\in\mathbb{N}} \xrightarrow{\|\cdot\|} f \quad \Rightarrow \quad (f_n(x))_{n\in\mathbb{N}} \to f(x) \quad \forall x.$$

- Example-1: For $\mathcal{H} := (C[0,1], \|\cdot\|_\infty)$

$$|f_n(x) - f(x)| \leqslant \max_{x\in[0,1]} |f_n(x) - f(x)| = \|f_n - f\|_\infty \xrightarrow{n\to\infty} 0,$$

- Def-4 intuition: We want

$$(f_n)_{n\in\mathbb{N}} \xrightarrow{\|\cdot\|} f \quad \Rightarrow \quad (f_n(x))_{n\in\mathbb{N}} \to f(x) \quad \forall x.$$

- Example-1: For $\mathcal{H} := (C[0,1], \|\cdot\|_\infty)$

$$|f_n(x) - f(x)| \leqslant \max_{x\in[0,1]} |f_n(x) - f(x)| = \|f_n - f\|_\infty \xrightarrow{n\to\infty} 0,$$

but no inner product in $C[0,1]$ (as we saw it – parallelograms).

Kernels: Definition-4 – continued

Let us now try a Hilbert space: $\mathcal{H} = L^2[0,1] \ni f_n(x) = x^n$ (simple).

Kernels: Definition-4 – continued

Let us now try a Hilbert space: $\mathcal{H} = L^2[0,1] \ni f_n(x) = x^n$ (simple).

1. $f_n \xrightarrow{n\to\infty} 0 := f^* \in \mathcal{H}$ since

$$\lim_{n\to\infty} \|f_n - 0\|_{L^2[0,1]} = \lim_{n\to\infty} \left(\int_0^1 x^{2n} \mathrm{d}x\right)^{1/2}$$

Kernels: Definition-4 – continued

Let us now try a Hilbert space: $\mathcal{H} = L^2[0,1] \ni f_n(x) = x^n$ (simple).

1. $f_n \xrightarrow{n\to\infty} 0 := f^* \in \mathcal{H}$ since

$$\lim_{n\to\infty} \|f_n - 0\|_{L^2[0,1]} = \lim_{n\to\infty} \left(\int_0^1 x^{2n} \mathrm{d}x\right)^{1/2} = \lim_{n\to\infty} \sqrt{\left[\frac{x^{2n+1}}{2n+1}\right]_{x=0}^{x=1}}$$

Kernels: Definition-4 – continued

Let us now try a Hilbert space: $\mathcal{H} = L^2[0,1] \ni f_n(x) = x^n$ (simple).

1. $f_n \xrightarrow{n\to\infty} 0 := f^* \in \mathcal{H}$ since

$$\lim_{n\to\infty} \|f_n - 0\|_{L^2[0,1]} = \lim_{n\to\infty} \left(\int_0^1 x^{2n} \mathrm{d}x\right)^{1/2} = \lim_{n\to\infty} \sqrt{\left[\frac{x^{2n+1}}{2n+1}\right]_{x=0}^{x=1}}$$
$$= \lim_{n\to\infty} \frac{1}{\sqrt{2n+1}} = 0,$$

Kernels: Definition-4 – continued

Let us now try a Hilbert space: $\mathcal{H} = L^2[0,1] \ni f_n(x) = x^n$ (simple).

1. $f_n \xrightarrow{n\to\infty} 0 := f^* \in \mathcal{H}$ since

$$\lim_{n\to\infty} \|f_n - 0\|_{L^2[0,1]} = \lim_{n\to\infty} \left(\int_0^1 x^{2n} \mathrm{d}x\right)^{1/2} = \lim_{n\to\infty} \sqrt{\left[\frac{x^{2n+1}}{2n+1}\right]_{x=0}^{x=1}}$$
$$= \lim_{n\to\infty} \frac{1}{\sqrt{2n+1}} = 0,$$

2. but $f_n(1) = 1 \nrightarrow f^*(1) = 0$.

In L^2: norm convergence $\not\Rightarrow$ pointwise convergence.

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g)$$

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f + g) = (f + g)(x)$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x)$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g),$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g),$$
$$\delta_x(\lambda f)$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g),$$
$$\delta_x(\lambda f) = (\lambda f)(x)$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g),$$
$$\delta_x(\lambda f) = (\lambda f)(x) = \lambda f(x)$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g),$$
$$\delta_x(\lambda f) = (\lambda f)(x) = \lambda f(x) = \lambda \delta_x(f).$$

Kernels: Definition-4

- Evaluation functional: $\delta_x(f) := f(x)$ is linear

$$\delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g),$$
$$\delta_x(\lambda f) = (\lambda f)(x) = \lambda f(x) = \lambda \delta_x(f).$$

- Def-4 (evaluation point of view): $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ Hilbert space,

$$\delta_x : f \in \mathcal{H} \mapsto f(x) \in \mathbb{R}$$

is continuous for all $x \in \mathcal{X}$.

Relation of Definition 1-4

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel, constructive):

$$k(\cdot, b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)] \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Relation of Definition 1-4

- Def-1 (feature space):

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel, constructive):

$$k(\cdot,b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot,b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)] \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

- All these definitions are equivalent, $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k$.
- Trickiest direction (Moore-Aronszajn theorem):

$$k \text{ positive definite function} \xrightarrow{\text{construction}} \text{RKHS}.$$

Example: every kernel is positive definite

$$\mathbf{a}^T \mathbf{G} \mathbf{a} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j)$$

Example: every kernel is positive definite

$$\mathbf{a}^T\mathbf{G}\mathbf{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) \overset{(i)}{=} \left\langle \sum_{i=1}^{n} a_i \varphi(x_i), \sum_{j=1}^{n} a_j \varphi(x_j) \right\rangle_{\mathcal{H}}$$

(i): k definition, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ linear

Example: every kernel is positive definite

$$\mathbf{a}^T \mathbf{G} \mathbf{a} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \overset{(i)}{=} \left\langle \sum_{i=1}^{n} a_i \varphi(x_i), \sum_{j=1}^{n} a_j \varphi(x_j) \right\rangle_{\mathcal{H}}$$

$$\overset{(ii)}{=} \left\| \sum_{i=1}^{n} a_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geqslant 0.$$

(i): k definition, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ linear, (ii) $\|\cdot\|_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$.

Kernels: further examples

- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$: $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$

$$k_p(\mathbf{x},\mathbf{y}) = (\langle \mathbf{x},\mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x},\mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

$$k_e(\mathbf{x},\mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}, \qquad k_C(\mathbf{x},\mathbf{y}) = 1 + \frac{1}{\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}.$$

Kernels: further examples

- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$: $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$

$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

$$k_e(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}, \qquad k_C(\mathbf{x}, \mathbf{y}) = 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}.$$

- $\mathcal{X}$ = strings:
  - r-spectrum kernel: # of common $\leqslant r$-substrings.

Kernels: further examples

- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$: $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$

$$k_p(\mathbf{x},\mathbf{y}) = (\langle \mathbf{x},\mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x},\mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

$$k_e(\mathbf{x},\mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}, \qquad k_C(\mathbf{x},\mathbf{y}) = 1 + \frac{1}{\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}.$$

- $\mathcal{X}$ = strings:
  - r-spectrum kernel: # of common $\leqslant r$-substrings.
- $\mathcal{X}$ = time-series: dynamic time-warping.

Kernel examples – continued

Matérn kernel: flexible family, well-suited for approximation (RFF)

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

where

- K_{ν}: modified Bessel function of the second kind of order ν,
- $\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x}\mathrm{d}x$: Gamma function ($t > 0$).

Kernel examples – continued

Matérn kernel: flexible family, well-suited for approximation (RFF)

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

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

where

- K_{ν}: modified Bessel function of the second kind of order ν,
- $\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x}\mathrm{d}x$: Gamma function ($t>0$).

Kernel examples – continued

Matérn kernel: flexible family, well-suited for approximation (RFF)

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

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

where

- K_{ν}: modified Bessel function of the second kind of order ν,
- $\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x}\mathrm{d}x$: Gamma function ($t>0$).

Specific cases:

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

Kernel puzzle

Let

$$\mathcal{X} = \{0, 1\},$$
$$k(x, x') = \left\{ \begin{array}{ll} 1, & \text{if } x \neq x' \\ -1, & \text{if } x = x' \end{array} \right\}.$$

Kernel puzzle

Let

$$\mathcal{X} = \{0, 1\},$$
$$k(x, x') = \left\{ \begin{array}{ll} 1, & \text{if } x \neq x' \\ -1, & \text{if } x = x' \end{array} \right\}.$$

Puzzle

Is k a kernel?

Kernel puzzle

Let

$$\mathcal{X} = \{0, 1\},$$

$$k(x, x') = \left\{ \begin{array}{ll} 1, & \text{if } x \neq x' \\ -1, & \text{if } x = x' \end{array} \right\}.$$

Puzzle

Is k a kernel?

No!

$$k(x, x) = \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}}$$

Kernel puzzle

Let

$$\mathcal{X} = \{0,1\},$$
$$k(x,x') = \left\{ \begin{array}{ll} 1, & \text{if } x \neq x' \\ -1, & \text{if } x = x' \end{array} \right\}.$$

Puzzle

Is k a kernel?

No!

$$k(x,x) = \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}} = \|\varphi(x)\|_{\mathcal{H}}^2 \geqslant 0 \quad \text{(Gram with } n = 1\text{)},$$

Kernel puzzle

Let

$$\mathcal{X} = \{0, 1\},$$
$$k(x, x') = \left\{ \begin{array}{ll} 1, & \text{if } x \neq x' \\ -1, & \text{if } x = x' \end{array} \right\}.$$

Puzzle

Is k a kernel?

No!

$$k(x, x) = \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}} = \|\varphi(x)\|_{\mathcal{H}}^2 \geqslant 0 \quad \text{(Gram with } n = 1\text{)},$$
$$k(0, 0) = k(1, 1) = -1 \quad \text{(in our case)}.$$

Kernel puzzle

Let

$$\mathcal{X} = \{0,1\},$$
$$k(x,x') = \begin{Bmatrix} 1, & \text{if } x \neq x' \\ -1, & \text{if } x = x' \end{Bmatrix}.$$

Puzzle

Is k a kernel?

No!

$$k(x,x) = \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}} = \|\varphi(x)\|_{\mathcal{H}}^2 \geqslant 0 \quad \text{(Gram with } n = 1\text{)},$$
$$k(0,0) = k(1,1) = -1 \quad \text{(in our case)}.$$

Easy-to-check conditions for a $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ function to be kernel ?

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ is a kernel.

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why?

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow$ $k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why? $\Leftarrow$ Gram.

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geqslant 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geqslant 0$ $(m = 1, \ldots, M)$, then

$$\sum_{m=1}^{M} \alpha_m k_m(x, y) = \; ?$$

Kernel factory

- We know: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow$ $k+\gamma$: kernel ($\gamma \in \mathbb{R}^{\geqslant 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geqslant 0$ ($m = 1,\ldots,M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x,y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y)\rangle_{\mathcal{H}_m}$$

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geqslant 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geqslant 0$ ($m = 1, \ldots, M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x, y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y) \rangle_{\mathcal{H}_m}$$
$$= \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}},$$

Kernel factory

- We know: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow$ $k+\gamma$: kernel ($\gamma \in \mathbb{R}^{\geqslant 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geqslant 0$ ($m = 1,\ldots,M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x,y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y)\rangle_{\mathcal{H}_m}$$
$$= \langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}},$$
$$\varphi(x) = \left(\sqrt{\alpha_1}\varphi_1(x),\ldots,\sqrt{\alpha_M}\varphi_M(x)\right) \in \mathcal{H} := \oplus_{m=1}^{M}\mathcal{H}_m.$$

Kernel factory

- We know: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k+\gamma$: kernel ($\gamma \in \mathbb{R}^{\geqslant 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geqslant 0$ ($m = 1,\ldots,M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x,y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y)\rangle_{\mathcal{H}_m}$$
$$= \langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}},$$
$$\varphi(x) = (\sqrt{\alpha_1}\varphi_1(x),\ldots,\sqrt{\alpha_M}\varphi_M(x)) \in \mathcal{H} := \oplus_{m=1}^{M}\mathcal{H}_m.$$

Example: $\oplus_{m=1}^{M}\mathbb{R} = \mathbb{R}^M$.

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^M$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^M k_m\right)\left(\left(x_1,\ldots,x_M\right),\left(x_1',\ldots,x_M'\right)\right)=\prod_{m=1}^M k_m\left(x_m,x_m'\right).$$

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^M$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^M k_m\right)\left(\left(x_1,\ldots,x_M\right),\left(x_1',\ldots,x_M'\right)\right)=\prod_{m=1}^M k_m\left(x_m,x_m'\right).$$

- Thus, $(k_m)_{m=1}^M : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ kernels $\Rightarrow \prod_{m=1}^M k_m(x,x')$: kernel on $\mathcal{X}$.

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^M$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^M k_m\right)\left(\left(x_1,\ldots,x_M\right),\left(x_1',\ldots,x_M'\right)\right)=\prod_{m=1}^M k_m\left(x_m,x_m'\right).$$

- Thus, $(k_m)_{m=1}^M : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ kernels $\Rightarrow \prod_{m=1}^M k_m(x,x')$: kernel on $\mathcal{X}$.
- Recall: $\otimes_{m=1}^M k_m$ will be in HSIC.

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^M$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^M k_m\right)\left(\left(x_1,\ldots,x_M\right),\left(x_1',\ldots,x_M'\right)\right)=\prod_{m=1}^M k_m\left(x_m,x_m'\right).$$

- Thus, $(k_m)_{m=1}^M : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ kernels $\Rightarrow \prod_{m=1}^M k_m(x,x')$: kernel on $\mathcal{X}$.
- Recall: $\otimes_{m=1}^M k_m$ will be in HSIC.
- Consequence ($\gamma\geqslant 0$, $p\in\mathbb{Z}^+$):

$$k(\mathbf{x},\mathbf{y})=\left(\langle\mathbf{x},\mathbf{y}\rangle_2+\gamma\right)^p$$

is a kernel.

Kernel factory: product indeed

Intuition for $M = 2$ and assuming $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$(k_1 \otimes k_2)\left((x, y), (x', y')\right) = k_1(x, x')k_2(y, y')$$

Intuition for $M = 2$ and assuming $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$
\begin{aligned}
(k_1 \otimes k_2)\left((x, y), (x', y')\right) &= k_1(x, x') k_2(y, y') \\
&= \left\langle \varphi_1(x), \varphi_1(x') \right\rangle_{\mathcal{H}_1} \left\langle \varphi_2(y), \varphi_2(y') \right\rangle_{\mathcal{H}_2}
\end{aligned}
$$

Intuition for $M = 2$ and assuming $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$\begin{aligned}(k_1 \otimes k_2)\left((x,y),(x',y')\right) &= k_1(x,x')k_2(y,y')\\ &= \left\langle \varphi_1(x), \varphi_1(x')\right\rangle_{\mathcal{H}_1} \left\langle \varphi_2(y), \varphi_2(y')\right\rangle_{\mathcal{H}_2}\\ &= \varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y)\end{aligned}$$

Intuition for $M = 2$ and assuming $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$
\begin{aligned}
(k_1 \otimes k_2)\left((x,y),(x',y')\right) &= k_1(x,x')k_2(y,y') \\
&= \left\langle \varphi_1(x), \varphi_1(x') \right\rangle_{\mathcal{H}_1} \left\langle \varphi_2(y), \varphi_2(y') \right\rangle_{\mathcal{H}_2} \\
&= \varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y) \\
&= \operatorname{tr}\left(\varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y)\right)
\end{aligned}
$$

Kernel factory: product indeed

Intuition for $M = 2$ and assuming $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$\begin{aligned}
(k_1 \otimes k_2)\left((x,y),(x',y')\right) &= k_1(x,x')k_2(y,y') \\
&= \left\langle \varphi_1(x), \varphi_1(x')\right\rangle_{\mathcal{H}_1} \left\langle \varphi_2(y), \varphi_2(y')\right\rangle_{\mathcal{H}_2} \\
&= \varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y) \\
&= \operatorname{tr}\left(\varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y)\right) \\
&= \operatorname{tr}\left(\varphi_2(y)\varphi_1(x)^T \varphi_1(x') \varphi_2(y')^T\right)
\end{aligned}$$

Kernel factory: product indeed

Intuition for $M = 2$ and assuming $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$\begin{aligned}
(k_1 \otimes k_2)\left((x,y),(x',y')\right) &= k_1(x,x')k_2(y,y') \\
&= \left\langle \varphi_1(x), \varphi_1(x') \right\rangle_{\mathcal{H}_1} \left\langle \varphi_2(y), \varphi_2(y') \right\rangle_{\mathcal{H}_2} \\
&= \varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y) \\
&= \operatorname{tr}\left(\varphi_1(x)^T \varphi_1(x') \varphi_2\left(y'\right)^T \varphi_2(y)\right) \\
&= \operatorname{tr}\left(\varphi_2(y)\varphi_1(x)^T \varphi_1(x') \varphi_2(y')^T\right) \\
&= \left\langle \underbrace{\varphi_1(x)\varphi_2(y)^T}_{\in \mathbb{R}^{d_1 \times d_2}}, \underbrace{\varphi_1(x')\varphi_2(y')^T}_{\in \mathbb{R}^{d_1 \times d_2}} \right\rangle_F,
\end{aligned}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle_F = \operatorname{tr}\left(\mathbf{A}^T \mathbf{B}\right) = \sqrt{\sum_{ij} A_{ij} B_{ij}}$ is the Frobenius inner product.

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why?

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why? $\Leftarrow$ Gram.

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why? $\Leftarrow$ Gram.
Example ($\gamma > 0$):

$$k(\mathbf{x},\mathbf{y}) = e^{\gamma\langle \mathbf{x},\mathbf{y}\rangle_2} = \sum_{n\in\mathbb{N}} \frac{\left(\gamma\langle \mathbf{x},\mathbf{y}\rangle_2\right)^n}{n!}$$

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why? $\Leftarrow$ Gram.
Example ($\gamma > 0$):

$$k(\mathbf{x},\mathbf{y}) = e^{\gamma\langle\mathbf{x},\mathbf{y}\rangle_2} = \sum_{n\in\mathbb{N}} \frac{\left(\gamma\langle\mathbf{x},\mathbf{y}\rangle_2\right)^n}{n!}$$

Reason: polynomial kernel & limit rule.

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x, y) = f(x)k(x, y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x, y) = f(x)k(x, y)f(y)$$

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f:\mathcal{X}\to\mathbb{R}$, then

$$\tilde{k}(x,y)=f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x,y)=f(x)k(x,y)f(y)=f(x)\langle\varphi(x),\varphi(y)\rangle_{\mathcal{H}}f(y)$$

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x,y) = f(x)k(x,y)f(y) = f(x)\langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}} f(y)$$
$$= \left\langle \underbrace{f(x)\varphi(x)}_{=:\tilde{\varphi}(x)}, f(y)\varphi(y) \right\rangle_{\mathcal{H}}.$$

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f:\mathcal{X}\to\mathbb{R}$, then

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x,y) = f(x)k(x,y)f(y) = f(x)\langle\varphi(x),\varphi(y)\rangle_{\mathcal{H}}f(y)$$
$$= \Big\langle \underbrace{f(x)\varphi(x)}_{=:\tilde{\varphi}(x)}, f(y)\varphi(y)\Big\rangle_{\mathcal{H}}.$$

Example (Gaussian kernel, $\gamma > 0$): previous example & new rule

$$k(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}$$

by using $\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\langle\mathbf{x},\mathbf{y}\rangle$.

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x,y) = f(x)k(x,y)f(y) = f(x)\langle\varphi(x),\varphi(y)\rangle_{\mathcal{H}}f(y)$$
$$= \Big\langle \underbrace{f(x)\varphi(x)}_{=:\tilde{\varphi}(x)}, f(y)\varphi(y) \Big\rangle_{\mathcal{H}}.$$

Example (Gaussian kernel, $\gamma > 0$): previous example & new rule

$$k(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2} = e^{-\gamma\|\mathbf{x}\|_2^2} e^{2\gamma\langle\mathbf{x},\mathbf{y}\rangle_2} e^{-\gamma\|\mathbf{x}\|_2^2}$$

by using $\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\langle\mathbf{x},\mathbf{y}\rangle$.

Kernel factory: $\mathbb{R}^d$ & Bochner theorem

We focus on continuous bounded shift-invariant kernels:

Theorem (Bochner's theorem [Wendland, 2005])

$$k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y}) = \int_{\mathbb{R}^d} e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle} d\Lambda(\boldsymbol{\omega}),$$

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

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

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

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

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

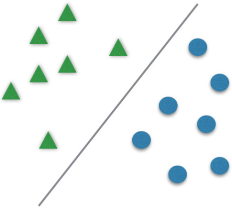
For Poisson kernel: $\sigma \in (0,1)$.

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

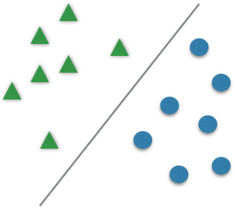
For $\mathbf{x} \in \mathbb{R}^d$: $k_0(\mathbf{x}) = \prod_{j=1}^{d} k_0(x_j)$, $\widehat{k_0}(\boldsymbol{\omega}) = \prod_{j=1}^{d}\widehat{k_0}(\omega_j)$.

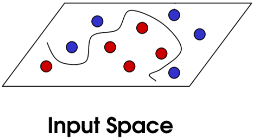
Kernels in action: classification, regression, dimensionality reduction

- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.



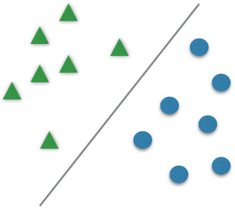
- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.

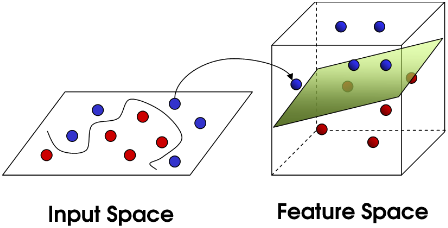




Classification, regression

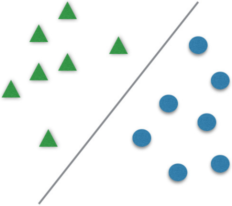
- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.

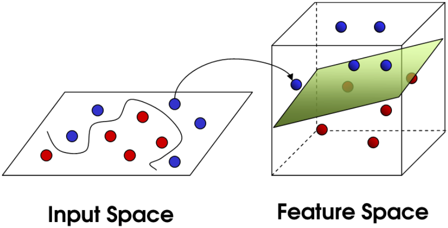




Classification, regression

- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.
- Regression similarly: $y_i \in \mathbb{R}$.





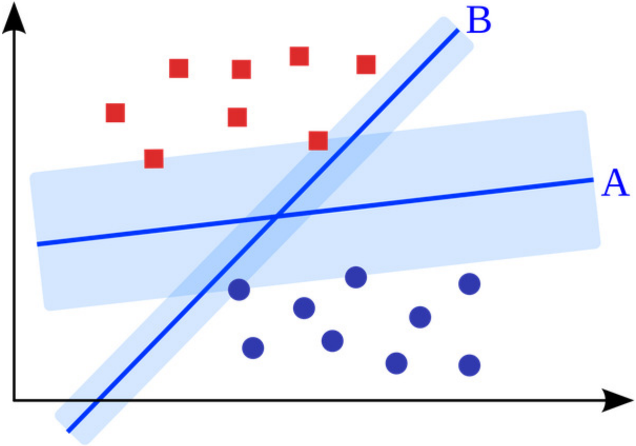
Dimensionality reduction: intuition

- Given: a set of observations $X = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^D$.
- Goal: find $X' = \{\mathbf{x}_i'\}_{i=1}^n \subset \mathbb{R}^d$ 'preserving' the geometry of X.
- $d \ll D$: compression (images, music, ...).

Classification: SVM

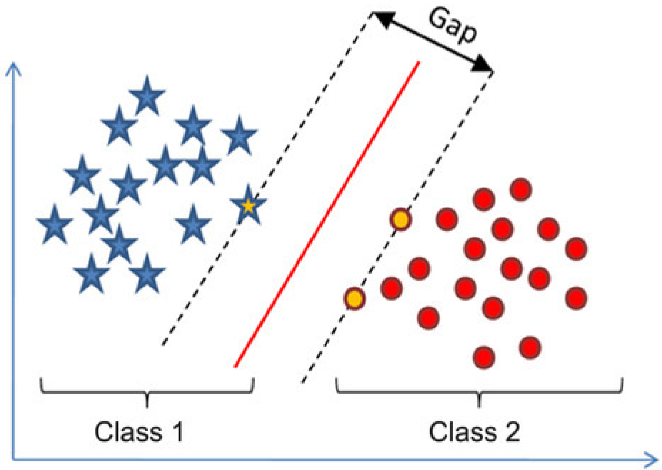
Support vector machine (SVM) for classification

Which separating line is the 'best'?



Support Vector Machine (SVM)

SVM answer: the one with the largest margin.



SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.
  - $\mathbf{w}$: normal vector, b: offset.

SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.
  - $\mathbf{w}$: normal vector, b: offset.
- Goal:

$$\max_{\mathbf{w},b} \underbrace{\frac{2}{\|\mathbf{w}\|_2}}_{\text{margin}} \Leftrightarrow \min \|\mathbf{w}\|_2^2, \text{ s.t. } \underbrace{\begin{cases} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geqslant 1 & \text{if } y_i = 1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leqslant -1 & \text{otherwise.} \end{cases}}_{\text{correction classification}}$$

SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.
  - $\mathbf{w}$: normal vector, b: offset.
- Goal:

$$\max_{\mathbf{w},b} \underbrace{\frac{2}{\|\mathbf{w}\|_2}}_{\text{margin}} \Leftrightarrow \min \|\mathbf{w}\|_2^2, \text{ s.t. } \underbrace{\begin{cases} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geqslant 1 & \text{if } y_i = 1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leqslant -1 & \text{otherwise.} \end{cases}}_{\text{correction classification}}$$

- Shortly,

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2, \text{ s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geqslant 1, \forall i.$$

SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.
  - $\mathbf{w}$: normal vector, b: offset.
- Goal:

$$\max_{\mathbf{w},b} \underbrace{\frac{2}{\|\mathbf{w}\|_2}}_{\text{margin}} \Leftrightarrow \min \|\mathbf{w}\|_2^2, \text{ s.t. } \underbrace{\begin{cases} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geqslant 1 & \text{if } y_i = 1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leqslant -1 & \text{otherwise.} \end{cases}}_{\text{correction classification}}$$

- Shortly,

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2, \text{ s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geqslant 1, \forall i.$$

- Decision: $\hat{y}(\mathbf{x}) = \text{sign}\left(\langle \mathbf{w}, \mathbf{x} \rangle + b \right)$.

SVM formulation: soft classification

- Hard classification objective:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2 \;\; \text{s.t.} \;\; y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geqslant 1, \forall i.$$

There might not be solution! (non-linearly separable case)

SVM formulation: soft classification

- Hard classification objective:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2 \ \text{ s.t. } \ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geqslant 1, \forall i.$$

  There might not be solution! (non-linearly separable case)

- Soft classification objective:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i \ \text{ s.t. } \ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geqslant 1 - \xi_i, \ \xi_i \geqslant 0, \ \forall i.$$

  Linear penalty on misclassification.

SVM formulation: soft classification

Soft classification objective:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i, \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0 \quad (\forall i).$$

Lagrangian function: with $\alpha_i \geqslant 0, \beta_i \geqslant 0$ $(\forall i)$

$$L(\mathbf{w}, b, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{objective - Lagrangian multipliers} \times \text{conditions}$$
$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\beta_i\xi_i.$$

Solving for $\frac{\partial L}{\partial \text{primal}} = 0$, we get ...

SVM formulation: soft classification

$$L(\mathbf{w}, b, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) =$$
$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\beta_i\xi_i.$$

Optimality equations:

$$\mathbf{0} = \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \quad (\mathbf{w} \leftrightarrow \boldsymbol{\alpha}),$$
$$0 = \frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i,$$
$$0 = \frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i.$$

Plugging these equations back to L, we have ...

SVM formulation: soft classification

Dual form:

$$\max_{\boldsymbol{\alpha}} \underbrace{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j}_{\text{quadratic in } \boldsymbol{\alpha}}, \text{ s.t. } \underbrace{0 \leqslant \alpha_i \leqslant C, \sum_{i=1}^{n} \alpha_i y_i = 0}_{\text{linear in } \boldsymbol{\alpha}}.$$

SVM formulation: soft classification

Dual form:

$$\max_{\boldsymbol{\alpha}} \underbrace{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j}_{\text{quadratic in } \boldsymbol{\alpha}}, \text{ s.t. } \underbrace{0 \leqslant \alpha_i \leqslant C, \sum_{i=1}^{n} \alpha_i y_i = 0}_{\text{linear in } \boldsymbol{\alpha}}.$$

- $b \Leftarrow y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 \Leftarrow \alpha_i > 0$ [complementary slackness].

SVM formulation: soft classification

Dual form:

$$\max_{\boldsymbol{\alpha}} \underbrace{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j}_{\text{quadratic in } \boldsymbol{\alpha}}, \text{ s.t. } \underbrace{0 \leqslant \alpha_i \leqslant C, \sum_{i=1}^{n} \alpha_i y_i = 0}_{\text{linear in } \boldsymbol{\alpha}}.$$

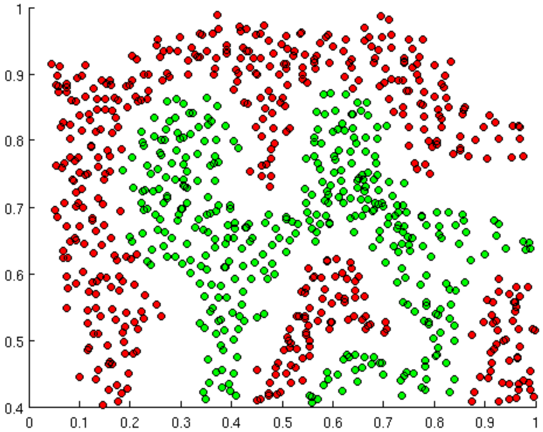
- $b \Leftarrow y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 \Leftarrow \alpha_i > 0$ [complementary slackness].
- QP: solvers are available.

If linear separability does not hold

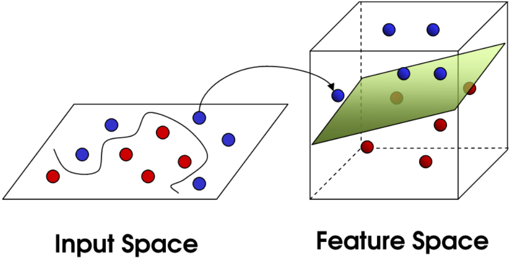
- Until this point:
    - (almost) linearly separable case.

If linear separability does not hold

- Until this point:
  - (almost) linearly separable case.
- Now:



If linear separability does not hold: kernel trick



Nonlinear SVM

- Linear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leqslant \alpha_i \leqslant C\ (\forall i).$$

Nonlinear SVM

- Linear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, \; 0 \leqslant \alpha_i \leqslant C \; (\forall i).$$

- Nonlinear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, \; 0 \leqslant \alpha_i \leqslant C \; (\forall i).$$

Nonlinear SVM

- Linear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leqslant \alpha_i \leqslant C\ (\forall i).$$

- Nonlinear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leqslant \alpha_i \leqslant C\ (\forall i).$$

- Nonlinear SVM (primal):

$$\min_{f \in \mathcal{H}_k, \boldsymbol{\xi}} \frac{1}{2} \|f\|_{\mathcal{H}_k}^2 + C \sum_{i=1}^{n} \xi_i, \text{ s.t. } y_i f(x_i) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ \forall i.$$

Kernel ridge regression

Kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} := \mathcal{H}_k$, $y_i \in \mathbb{R}$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

Kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} := \mathcal{H}_k$, $y_i \in \mathbb{R}$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

- Analytical solution (raised at distribution regression):

$$f(x) = [k(x_1, x), \ldots, k(x_n, x)]\,(\mathbf{G} + \lambda n I)^{-1}[y_1; \ldots; y_n],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n.$$

Kernel ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} := \mathcal{H}_k$, $y_i \in \mathbb{R}$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

- Analytical solution (raised at distribution regression):

$$f(x) = [k(x_1, x), \ldots, k(x_n, x)] (\mathbf{G} + \lambda n I)^{-1}[y_1; \ldots; y_n],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n.$$

Question

How do we get this solution?

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^T \mathbf{G}\mathbf{a} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{G}\mathbf{a} + \mathbf{a}^T\left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right]\mathbf{a}. \end{aligned}$$

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j)\rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^T\mathbf{G}\mathbf{a} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{G}\mathbf{a} + \mathbf{a}^T\left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right]\mathbf{a}. \end{aligned}$$

Solving $\mathbf{0} = \frac{\partial J}{\partial \mathbf{a}}$, one gets $\mathbf{a}^* = (\mathbf{G} + \lambda n \mathbf{I})^{-1}\mathbf{y}$

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^T\mathbf{G}\mathbf{a} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{G}\mathbf{a} + \mathbf{a}^T\left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right]\mathbf{a}. \end{aligned}$$

Solving $\mathbf{0} = \frac{\partial J}{\partial \mathbf{a}}$, one gets $\mathbf{a}^* = (\mathbf{G} + \lambda n \mathbf{I})^{-1}\mathbf{y}$ by

$$\frac{\partial \mathbf{a}^T\mathbf{B}\mathbf{a}}{\partial \mathbf{a}} = \left(\mathbf{B} + \mathbf{B}^T\right)\mathbf{a}, \quad \frac{\partial \mathbf{c}^T\mathbf{a}}{\partial \mathbf{a}} = \mathbf{c}.$$

Local Summary

- Motivation: infoT objectives, hypothesis testing.
- Kernels, RKHS: definitions, construction.
- Kernel applications: classification, ridge regression.

Notes

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y \in \mathcal{X}} k(x,y) \leqslant C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \leqslant C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\leqslant} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \leqslant C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\leqslant} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \leqslant C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\leqslant} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.
- k: bounded and continuous $\Rightarrow \forall f \in \mathcal{H}_k$ is bounded & continuous.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \leqslant C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\leqslant} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.
- k: bounded and continuous $\Rightarrow \forall f \in \mathcal{H}_k$ is bounded & continuous.
- $k \in C^m \Rightarrow \forall f \in \mathcal{H}_k$ is m-times continuously differentiable.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \leqslant C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\leqslant} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.
- k: bounded and continuous $\Rightarrow \forall f \in \mathcal{H}_k$ is bounded & continuous.
- $k \in C^m \Rightarrow \forall f \in \mathcal{H}_k$ is m-times continuously differentiable.
- k: analytic $\Rightarrow \forall f \in \mathcal{H}_k$ is analytic.

Hard vs soft-SVM classification

Recall:

- Hard SVM:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geqslant 1, \forall i.$$

- Soft SVM:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ \forall i$$

Hard vs soft-SVM classification

Recall:

- Hard SVM:

$$\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|_2^2, \text{ s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geqslant 1, \forall i.$$

- Soft SVM:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i \text{ s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ \forall i. \Leftrightarrow$$

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \max\left(1 - y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right), 0\right),$$

Hard vs soft-SVM classification

Recall:

- Hard SVM:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geqslant 1, \forall i.$$

- Soft SVM:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ \forall i. \Leftrightarrow$$

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \underbrace{\max\Big(1 - y_i\big(\underbrace{\langle \mathbf{w}, \mathbf{x}_i\rangle + b}_{=f(\mathbf{x}_i)}\big), 0\Big)}_{=:h(y_i f(\mathbf{x}_i))},$$

where $h(u) = \max(1-u, 0)$ is the hinge loss.

Hard vs soft-SVM classification – continued

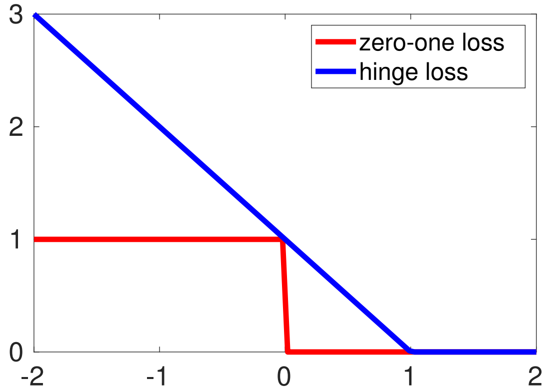
The hinge loss is the convex envelope of the zero-one loss:

$$z(u) = \mathbb{I}_{u<0}, \qquad u = y_i f(x_i),$$
$$h(u) = \max(1-u, 0).$$

Hard vs soft-SVM classification – continued

The hinge loss is the **convex envelope of the zero-one loss**:

$$z(u) = \mathbb{I}_{u<0}, \qquad u = y_i f(x_i),$$
$$h(u) = \max(1-u, 0).$$



Representer theorem [Schölkopf et al., 2001, Yu et al., 2013]

- Given: $\{(x_i, y_i)\}_{i=1}^n$, say classification/regression.
- Goal:

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{\mathcal{H}_k},$$

r : monotonically increasing.

Representer theorem [Schölkopf et al., 2001, Yu et al., 2013]

- Given: $\{(x_i, y_i)\}_{i=1}^n$, say classification/regression.
- Goal:

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{\mathcal{H}_k},$$

r : monotonically increasing.

- Example:

$$V(\ldots) = \frac{1}{n}\sum_{i=1}^n \max(1 - y_i f(x_i), 0) \quad \text{(soft classification)},$$

$$V(\ldots) = \frac{1}{n}\sum_{i=1}^n [f(x_i) - y_i]^2 \quad \text{(regression)}.$$

Representer theorem – continued

...then

- $\exists$ solution in the form:

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \quad \alpha_i \in \mathbb{R}.$$

- r: strictly increasing $\Rightarrow$ $\forall$ solution is of this form.
- Example: $r(z) = \lambda z$, $\lambda > 0$.

Representer theorem – proof

Objective

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|_{\mathcal{H}_k}^2\right) \to \min_{\mathcal{H}_k}.$$

Decompose & Pythagorean theorem:

$$\begin{aligned} S &= span\left(k(\cdot, x_i), i = 1, \ldots, n\right), \\ f &= f_S + f_\perp, \\ \|f\|_{\mathcal{H}_k}^2 &= \|f_S\|_{\mathcal{H}_k}^2 + \underbrace{\|f_\perp\|_{\mathcal{H}_k}^2}_{\geqslant 0} \geqslant \|f_S\|_{\mathcal{H}_k}^2. \end{aligned}$$

Representer theorem – proof

Objective

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{\mathcal{H}_k}.$$

Decompose & Pythagorean theorem:

$$\begin{aligned} S &= span\left(k(\cdot, x_i), i = 1, \ldots, n\right), \\ f &= f_S + f_\perp, \\ \|f\|^2_{\mathcal{H}_k} &= \|f_S\|^2_{\mathcal{H}_k} + \underbrace{\|f_\perp\|^2_{\mathcal{H}_k}}_{\geqslant 0} \geqslant \|f_S\|^2_{\mathcal{H}_k}. \end{aligned}$$

In J

- 1st term: depends on f_S only, $f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}$.

Representer theorem – proof

Objective

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|_{\mathcal{H}_k}^2\right) \to \min_{\mathcal{H}_k}.$$

Decompose & Pythagorean theorem:

$$\begin{aligned} S &= span\left(k(\cdot, x_i), i = 1, \ldots, n\right), \\ f &= f_S + f_\perp, \\ \|f\|_{\mathcal{H}_k}^2 &= \|f_S\|_{\mathcal{H}_k}^2 + \underbrace{\|f_\perp\|_{\mathcal{H}_k}^2}_{\geqslant 0} \geqslant \|f_S\|_{\mathcal{H}_k}^2. \end{aligned}$$

In J

- 1st term: depends on f_S only, $f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}$.
- 2nd term: can only decrease by neglecting $f_\perp$ ($r \nearrow$).

M-fold cross-validation [$\theta := (C, \sigma)$]:

1. Split data:
   - training set (X_{tr}, Y_{tr}): $X_{val,i}, Y_{val,i}$, $i = 1, \ldots, M$.
   - test set: X_{te}, Y_{te}.

Parameter selection $\rightarrow$ cost: C, kernel parameter: σ

M-fold **cross-validation** $[\theta := (C, \sigma)]$:

1. Split data:
   - training set (X_{tr}, Y_{tr}): $X_{val,i}, Y_{val,i}$, $i = 1, \ldots, M$.
   - test set: X_{te}, Y_{te}.
2. For fixed θ: evaluate the average error while
   - trained on: $X_{tr} \backslash X_{val,i}$, $Y_{tr} \backslash Y_{val,i}$,
   - tested on: $X_{val,i}$, $Y_{val,i}$.

Parameter selection → cost: C, kernel parameter: σ

M-fold cross-validation $[\theta := (C, \sigma)]$:

1. Split data:
   - training set (X_{tr}, Y_{tr}): $X_{val,i}, Y_{val,i}$, $i = 1, \ldots, M$.
   - test set: X_{te}, Y_{te}.
2. For fixed θ: evaluate the average error while
   - trained on: $X_{tr} \backslash X_{val,i}$, $Y_{tr} \backslash Y_{val,i}$,
   - tested on: $X_{val,i}$, $Y_{val,i}$.
3. $\theta^* :=$ minimizer of CV error.

Parameter selection $\rightarrow$ cost: C, kernel parameter: σ

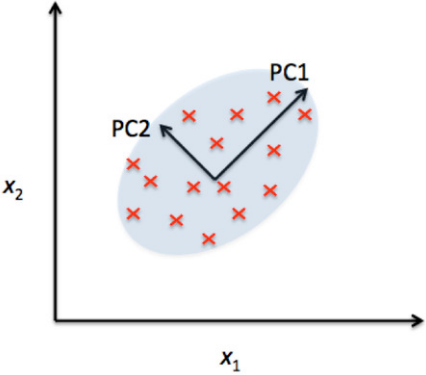
M-fold cross-validation $[\theta := (C, \sigma)]$:

1. Split data:
   - training set (X_{tr}, Y_{tr}): $X_{val,i}, Y_{val,i}$, $i = 1, \ldots, M$.
   - test set: X_{te}, Y_{te}.
2. For fixed θ: evaluate the average error while
   - trained on: $X_{tr} \backslash X_{val,i}$, $Y_{tr} \backslash Y_{val,i}$,
   - tested on: $X_{val,i}$, $Y_{val,i}$.
3. $\theta^* :=$ minimizer of CV error.
4. Report: performance of θ^* on X_{te}, Y_{te}.

PCA and its kernelized version

PCA: intuition

Task: find the best d-dimensional subspace approximating $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^D$.

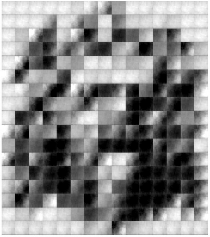


PCA example: 100%

(A)

PCA example: 100% $\longrightarrow$ 1%

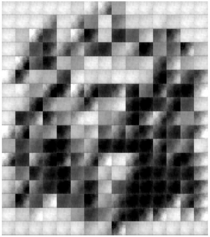
(A)



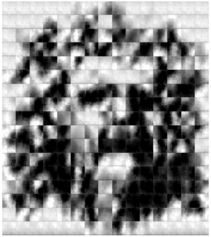
(B)

PCA example: 100% → 2%

(A)



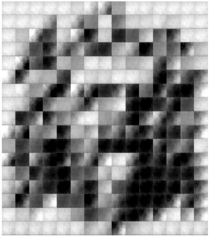
(B)



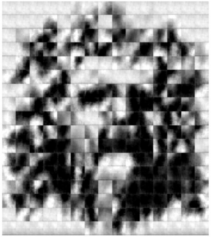
(C)

PCA example: 100% → 5%

(A)



(B)

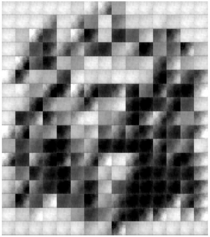


(C)

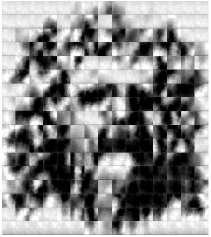
(D)

PCA example: 100% $\to$ 10%

(A)



(B)

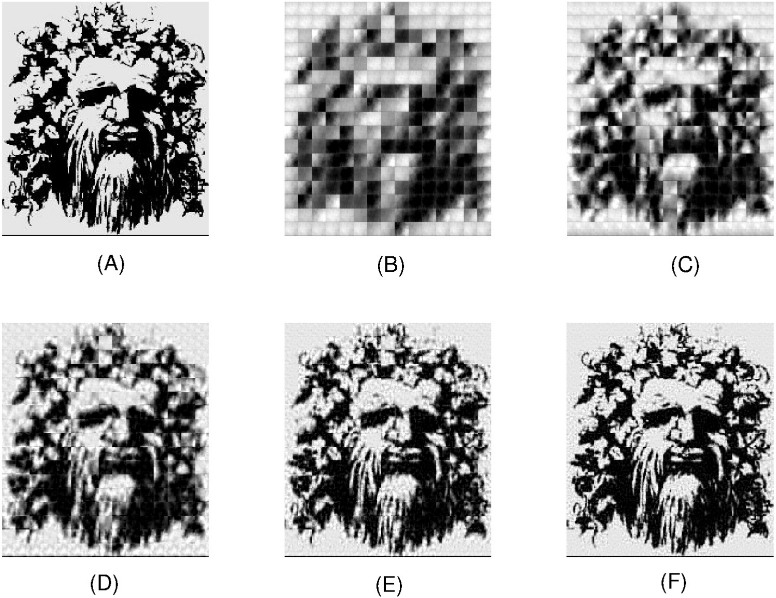


(C)

(D)

(E)

PCA example: 100% → 20%



PCA formulation: $d = 1$

- We are looking for the best one-dimensional projection.

- $\mathbb{E}$:= empirical/population expectation: $\mathbb{E}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$.
- Assumption: $\mathbb{E}\mathbf{x} = \mathbf{0}$.

PCA formulation: $d = 1$

- We are looking for the best one-dimensional projection.

- $\mathbb{E}$:= empirical/population expectation: $\mathbb{E}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$.
- Assumption: $\mathbb{E}\mathbf{x} = \mathbf{0}$.
  - centering: $\mathbf{x} \to \mathbf{x} - \mathbb{E}\mathbf{x}$.

PCA: projection

Projection ($\|\mathbf{w}\|_2 = 1$):

- $\hat{\mathbf{x}} = \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}$.
- zero mean: $\mathbf{0} \stackrel{?}{=} \mathbb{E}\hat{\mathbf{x}} = \mathbb{E}\left[\langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}\right]$

PCA: projection

Projection ($\|\mathbf{w}\|_2 = 1$):

- $\hat{\mathbf{x}} = \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}$.
- zero mean: $\mathbf{0} \stackrel{?}{=} \mathbb{E}\hat{\mathbf{x}} = \mathbb{E}\left[\langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}\right] = \Big\langle \mathbf{w}, \underbrace{\mathbb{E}\mathbf{x}}_{=\mathbf{0}} \Big\rangle \mathbf{w}$.

PCA: min residual $\Leftrightarrow$ max squared projection

- Goal: $\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}}$.

PCA: min residual $\Leftrightarrow$ max squared projection

- Goal: $\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}}$.
- Residual $\Rightarrow$ objective:

$$\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 = \left\|\mathbf{x}-\langle \mathbf{w},\mathbf{x}\rangle \mathbf{w}\right\|_2^2$$

PCA: min residual $\Leftrightarrow$ max squared projection

- Goal: $\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}}$.
- Residual $\Rightarrow$ objective:

$$\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 = \left\|\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w}\right\|_2^2 \overset{\|\mathbf{w}\|_2^2=1}{=} \left\|\mathbf{x}\right\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2 \Rightarrow$$

PCA: min residual ⇔ max squared projection

- Goal: $\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}}$.
- Residual ⇒ objective:

$$\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 = \left\|\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w}\right\|_2^2 \overset{\|\mathbf{w}\|_2^2=1}{=} \left\|\mathbf{x}\right\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2 \Rightarrow$$

$$\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 = \mathbb{E}\left[\left\|\mathbf{x}\right\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2\right] = \underbrace{\mathbb{E}\left\|\mathbf{x}\right\|_2^2}_{\text{independent of }\mathbf{w}} - \mathbb{E}\langle\mathbf{w},\mathbf{x}\rangle^2 \Leftrightarrow$$

PCA: min residual $\Leftrightarrow$ max squared projection

- Goal: $\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 \to \min_{\mathbf{w}}$.
- Residual $\Rightarrow$ objective:

$$\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 = \left\|\mathbf{x}-\langle\mathbf{w},\mathbf{x}\rangle\mathbf{w}\right\|_2^2 \overset{\|\mathbf{w}\|_2^2=1}{=} \left\|\mathbf{x}\right\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2 \Rightarrow$$

$$\mathbb{E}\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2 = \mathbb{E}\left[\left\|\mathbf{x}\right\|_2^2 - \langle\mathbf{w},\mathbf{x}\rangle^2\right] = \underbrace{\mathbb{E}\left\|\mathbf{x}\right\|_2^2}_{\text{independent of } \mathbf{w}} - \mathbb{E}\langle\mathbf{w},\mathbf{x}\rangle^2 \Leftrightarrow$$

Solution

maximizes the mean squared projection.

PCA: max squared projection $\Leftrightarrow$ max variance of projection

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + \mathrm{var}(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\langle \mathbf{w}, \mathbf{x} \rangle^2 = \Big(\underbrace{\mathbb{E}\langle \mathbf{w}, \mathbf{x} \rangle}_{=0} \Big)^2 + \mathrm{var}(\langle \mathbf{w}, \mathbf{x} \rangle).$$

PCA: max squared projection $\Leftrightarrow$ max variance of projection

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + \text{var}(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 = \big(\underbrace{\mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle}_{=0}\big)^2 + \text{var}(\left\langle \mathbf{w}, \mathbf{x} \right\rangle).$$

To sum up:

$$\text{Minimize MSE of the residual} : \min_{\mathbf{w}} \mathbb{E}\left\| \mathbf{x} - \hat{\mathbf{x}} \right\|_2^2 \Leftrightarrow$$

PCA: max squared projection ⇔ max variance of projection

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + \text{var}(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\langle \mathbf{w}, \mathbf{x} \rangle^2 = \Big(\underbrace{\mathbb{E}\langle \mathbf{w}, \mathbf{x} \rangle}_{=0} \Big)^2 + \text{var}(\langle \mathbf{w}, \mathbf{x} \rangle).$$

To sum up:

$$\text{Minimize MSE of the residual} : \min_{\mathbf{w}} \mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \Leftrightarrow$$

$$\text{Maximize mean squared projection} : \max_{\mathbf{w}} \mathbb{E}\langle \mathbf{w}, \mathbf{x} \rangle^2 \Leftrightarrow$$

By using $\mathbb{E}y^2 = (\mathbb{E}y)^2 + \mathrm{var}(y)$:

$$\max_{\mathbf{w}} \leftarrow \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 = \Big(\underbrace{\mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle}_{=0} \Big)^2 + \mathrm{var}(\left\langle \mathbf{w}, \mathbf{x} \right\rangle).$$

To sum up:

$$\begin{aligned} \text{Minimize MSE of the residual} &: \min_{\mathbf{w}} \mathbb{E}\left\| \mathbf{x} - \hat{\mathbf{x}} \right\|_2^2 \Leftrightarrow \\ \text{Maximize mean squared projection} &: \max_{\mathbf{w}} \mathbb{E}\left\langle \mathbf{w}, \mathbf{x} \right\rangle^2 \Leftrightarrow \\ \text{Maximize variance of the projection} &: \max_{\mathbf{w}} \mathrm{var}(\left\langle \mathbf{w}, \mathbf{x} \right\rangle). \end{aligned}$$

PCA: Optimization

By the bilinearity of cov:

$$\mathrm{var}\left(\langle \mathbf{w}, \mathbf{x} \rangle\right) = \mathrm{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right)$$

PCA: Optimization

By the bilinearity of cov:

$$\mathrm{var}\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) = \mathrm{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T \mathrm{cov}(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}.$$

PCA: Optimization

By the bilinearity of cov:

$$\mathrm{var}\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) = \mathrm{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T\,\mathrm{cov}(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}.$$

Lagrange function, solving for 'derivatives = 0':

PCA: Optimization

By the bilinearity of cov:

$$\operatorname{var}\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) = \operatorname{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T \operatorname{cov}(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}.$$

Lagrange function, solving for 'derivatives = 0':

$$L(\mathbf{w},\lambda) = \underbrace{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

PCA: Optimization

By the bilinearity of cov:

$$\mathrm{var}\left(\langle \mathbf{w}, \mathbf{x} \rangle\right) = \mathrm{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T \mathrm{cov}(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}.$$

Lagrange function, solving for 'derivatives = 0':

$$L(\mathbf{w}, \lambda) = \underbrace{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

$$0 = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda} = -(\mathbf{w}^T\mathbf{w} - 1),$$

PCA: Optimization

By the bilinearity of cov:

$$\text{var}\left(\langle \mathbf{w}, \mathbf{x} \rangle\right) = \text{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T \text{cov}(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}.$$

Lagrange function, solving for 'derivatives = 0':

$$L(\mathbf{w}, \lambda) = \underbrace{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

$$0 = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda} = -(\mathbf{w}^T\mathbf{w} - 1),$$

$$\mathbf{0} = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = 2\mathbf{\Sigma}\mathbf{w} - 2\lambda\mathbf{w} \Rightarrow$$

PCA: Optimization

By the bilinearity of cov:

$$\text{var}\left(\langle \mathbf{w}, \mathbf{x} \rangle\right) = \text{cov}\left(\mathbf{w}^T\mathbf{x}, \mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T \text{cov}(\mathbf{x})\mathbf{w} =: \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \to \max_{\|\mathbf{w}\|_2=1}.$$

Lagrange function, solving for 'derivatives $= 0$':

$$L(\mathbf{w}, \lambda) = \underbrace{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}_{=\text{objective}} - \lambda(\underbrace{\mathbf{w}^T\mathbf{w} - 1}_{=\text{condition}}) \Rightarrow$$

$$0 = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda} = -(\mathbf{w}^T\mathbf{w} - 1),$$

$$\mathbf{0} = \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = 2\mathbf{\Sigma}\mathbf{w} - 2\lambda\mathbf{w} \Rightarrow$$

Solution

$\mathbf{w}^*$: eigenvector associated to $\lambda_{\max}(\mathbf{\Sigma})$.

PCA: $d \geqslant 1$

PCA ($d \geqslant 1$): basis, approximation

- Goal: approximate with a d-dimensional subspace.
- ONB in the subspace ($\mathbf{W}^T\mathbf{W} = \mathbf{I}$):

$$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{D \times d},$$

- Approximation:

$$\hat{\mathbf{x}} = \sum_{i=1}^{d} \langle \mathbf{w}_i, \mathbf{x} \rangle \mathbf{w}_i = \mathbf{W}\mathbf{W}^T\mathbf{x}.$$

PCA ($d \geqslant 1$): min residual $\Leftrightarrow$ max squared projection

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2$$

PCA ($d \geqslant 1$): min residual $\Leftrightarrow$ max squared projection

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x}-\mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \mathbf{x}^T \underbrace{\left(\mathbf{I}-\mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I}-\mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x}$$

PCA ($d \geqslant 1$): min residual $\Leftrightarrow$ max squared projection

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x}$$

$$= \|\mathbf{x}\|_2^2 - \left\|\mathbf{W}^T\mathbf{x}\right\|_2^2,$$

PCA ($d \geqslant 1$): min residual $\Leftrightarrow$ max squared projection

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x}$$

$$= \|\mathbf{x}\|_2^2 - \left\|\mathbf{W}^T\mathbf{x}\right\|_2^2,$$

$$\mathbb{E}\,\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \underbrace{\mathbb{E}\,\|\mathbf{x}\|_2^2}_{\text{independent of } \mathbf{W}} - \mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2.$$

PCA ($d \geqslant 1$): min residual $\Leftrightarrow$ max squared projection

Using $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}\right\|_2^2 = \mathbf{x}^T \underbrace{\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right)}_{=\mathbf{I}-2\mathbf{W}\mathbf{W}^T+\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{I}-\mathbf{W}\mathbf{W}^T} \mathbf{x}$$

$$= \|\mathbf{x}\|_2^2 - \left\|\mathbf{W}^T\mathbf{x}\right\|_2^2,$$

$$\mathbb{E}\,\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \underbrace{\mathbb{E}\,\|\mathbf{x}\|_2^2}_{\text{independent of } \mathbf{W}} - \mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2.$$

Thus $\min_{\mathbf{W}} \mathbb{E}\,\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \Leftrightarrow \max_{\mathbf{W}} \mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2$.

PCA ($d \geqslant 1$): max squared projection $\Leftrightarrow$ max variance of projection

Let $\mathbf{y} = \mathbf{W}^T\mathbf{x}$:

$$\mathbb{E}\,\|\mathbf{y}\|_2^2 - \|\mathbb{E}\mathbf{y}\|_2^2 = \mathrm{var}(\mathbf{y})?$$

PCA ($d \geqslant 1$): max squared projection $\Leftrightarrow$ max variance of projection

Let $\mathbf{y} = \mathbf{W}^T\mathbf{x}$:

$$\mathbb{E}\,\|\mathbf{y}\|_2^2 - \|\mathbb{E}\mathbf{y}\|_2^2 = \mathrm{var}(\mathbf{y})?$$

$$= \mathbb{E}\left[\sum_i y_i^2\right] - \sum_i (\mathbb{E}y_i)^2 = \sum_i \mathrm{var}(y_i) \Rightarrow$$

PCA ($d \geqslant 1$): max squared projection $\Leftrightarrow$ max variance of projection

Let $\mathbf{y} = \mathbf{W}^T\mathbf{x}$:

$$\mathbb{E}\,\|\mathbf{y}\|_2^2 - \|\mathbb{E}\mathbf{y}\|_2^2 = \mathrm{var}(\mathbf{y})?$$

$$= \mathbb{E}\left[\sum_i y_i^2\right] - \sum_i (\mathbb{E}y_i)^2 = \sum_i \mathrm{var}(y_i) \Rightarrow$$

$$\mathbb{E}\left\|\mathbf{W}^T\mathbf{x}\right\|_2^2 - \Big\|\underbrace{\mathbb{E}[\mathbf{W}^T\mathbf{x}]}_{=\mathbf{W}^T\mathbb{E}\mathbf{x}=\mathbf{0}}\Big\|_2^2 = \sum_i \mathrm{var}\left(\left(\mathbf{W}^T\mathbf{x}\right)_i\right) \to \max_{\mathbf{W}}.$$

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}).$$

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}).$$

- $\boldsymbol{\Sigma}$: symmetric, positive semi-definite $\Rightarrow$ $\{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \mathbf{\Sigma} = \mathrm{cov}(\mathbf{x}).$$

- $\mathbf{\Sigma}$: symmetric, positive semi-definite $\Rightarrow$ $\{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.
- Variance decomposition: $\mathrm{cov}(\mathbf{x}) = \sum_{i=1}^D \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \mathbf{\Sigma} = \mathrm{cov}(\mathbf{x}).$$

- $\mathbf{\Sigma}$: symmetric, positive semi-definite $\Rightarrow$ $\{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.
- Variance decomposition: $\mathrm{cov}(\mathbf{x}) = \sum_{i=1}^D \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.
- Energy preserved using d components: $\sum_{i=1}^d \lambda_i \Rightarrow$

$$R^2 = R^2(d) := \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i} \in [0, 1].$$

PCA: $d \geqslant 1$

- The d principal components:

$$\{\mathbf{w}_i\}_{i=1}^d = \text{ top } d \text{ eigenvectors of } \mathbf{\Sigma} = \operatorname{cov}(\mathbf{x}).$$

- $\mathbf{\Sigma}$: symmetric, positive semi-definite $\Rightarrow$ $\{\mathbf{w}_i\}$: ONS, $\lambda_i \geqslant 0$.
- Variance decomposition: $\operatorname{cov}(\mathbf{x}) = \sum_{i=1}^D \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.
- Energy preserved using d components: $\sum_{i=1}^d \lambda_i \Rightarrow$

$$R^2 = R^2(d) := \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i} \in [0,1].$$

- In practice: choose d such that $R^2 \approx 0.8 - 0.9$.

Kernel PCA: idea for '$d = 1$' $\leftrightarrow f$

Let $\mathcal{H} = \mathcal{H}_k$.

- Objective function:

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}\left\langle f, \underbrace{\varphi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)}_{=:\tilde{\varphi}(x_i)}\right\rangle^2 = \mathrm{var}(f) \to \max_{f:\|f\|_{\mathcal{H}}\leqslant 1}.$$

Kernel PCA: idea for '$d = 1$' $\leftrightarrow f$

Let $\mathcal{H} = \mathcal{H}_k$.

- Objective function:

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}\Big\langle f, \underbrace{\varphi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)}_{=:\tilde{\varphi}(x_i)}\Big\rangle^2 = \mathrm{var}(f) \to \max_{f:\|f\|_{\mathcal{H}}\leqslant 1}.$$

- The solution can be searched in the form ($\mathcal{H} \ni f \leftrightarrow \mathbf{a} \in \mathbb{R}^n$):

$$f = \sum_{i=1}^{n} a_i \tilde{\varphi}(x_i)$$

since component $\perp$ *span* $\left(\{\tilde{\varphi}(x_i)\}_{i=1}^n\right)$ has no contribution.

Kernel PCA: idea for '$d = 1$' $\leftrightarrow f$

Let $\mathcal{H} = \mathcal{H}_k$.

- Objective function:

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}\left\langle f, \underbrace{\varphi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)}_{=:\tilde{\varphi}(x_i)}\right\rangle^2 = \mathrm{var}(f) \to \max_{f:\|f\|_{\mathcal{H}}\leqslant 1}.$$

- The solution can be searched in the form ($\mathcal{H} \ni f \leftrightarrow \mathbf{a} \in \mathbb{R}^n$):

$$f = \sum_{i=1}^{n} a_i \tilde{\varphi}(x_i)$$

since component $\perp$ *span* $\left(\{\tilde{\varphi}(x_i)\}_{i=1}^n\right)$ has no contribution.

- We will get an eigenvalue problem for $\mathbf{a}$.

(Empirical) covariance operator

$$C := \frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}(x_i) \otimes \tilde{\varphi}(x_i).$$

$c \otimes d$ is the analogue of cd^T:

$$(c \otimes d)(e) = c \langle d, e \rangle_{\mathcal{H}}.$$

(Empirical) covariance operator

$$C := \frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}(x_i) \otimes \tilde{\varphi}(x_i).$$

$c \otimes d$ is the analogue of cd^T:

$$(c \otimes d)(e) = c \langle d, e \rangle_{\mathcal{H}}.$$

Similarly to the finite-dimensional case:

$$Cf_j = \lambda_j f_j.$$

Challenge

How do we solve this eigenvalue problem?

Computation of Cf_j

Assume j is fixed ($Cf = \lambda f$):

$$Cf = \left[\frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}(x_i) \otimes \tilde{\varphi}(x_i)\right] f$$

Computation of Cf_j

Assume j is fixed ($Cf = \lambda f$):

$$Cf = \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\otimes\tilde{\varphi}(x_i)\right] f$$

$$\stackrel{\otimes\text{ def}}{=} \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\left\langle\tilde{\varphi}(x_i),\sum_{j=1}^{n}a_j\tilde{\varphi}(x_j)\right\rangle_{\mathcal{H}}$$

Computation of Cf_j

Assume j is fixed ($Cf = \lambda f$):

$$Cf = \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\otimes\tilde{\varphi}(x_i)\right] f$$

$$\stackrel{\otimes \text{ def}}{=} \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\left\langle\tilde{\varphi}(x_i), \sum_{j=1}^{n} a_j\tilde{\varphi}(x_j)\right\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\sum_{j=1}^{n} a_j\tilde{k}(x_i, x_j)$$

with $\tilde{\mathbf{G}} = \mathbf{HGH} = \left[\tilde{k}(x_i, x_j)\right]_{i,j=1}^{n}$, $\mathbf{H} = \mathbf{I}_n - \frac{\mathbf{E}_n}{n}$.

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}(x_i) \sum_{j=1}^{n} a_j \tilde{k}(x_i, x_j)$.
- Idea: multiple by $\tilde{\varphi}(x_r)$

$$\langle \tilde{\varphi}(x_r), \lambda f \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \lambda \sum_{j=1}^{n} a_j \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}}$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\sum_{j=1}^{n}a_j\tilde{k}(x_i,x_j)$.
- Idea: multiple by $\tilde{\varphi}(x_r)$

$$\langle\tilde{\varphi}(x_r),\lambda f\rangle_{\mathcal{H}} = \left\langle\tilde{\varphi}(x_r),\lambda\sum_{j=1}^{n}a_j\tilde{\varphi}(x_j)\right\rangle_{\mathcal{H}} = \lambda\underbrace{\sum_{j=1}^{n}a_j\tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_r},$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}(x_i) \sum_{j=1}^{n} a_j \tilde{k}(x_i, x_j)$.
- Idea: multiple by $\tilde{\varphi}(x_r)$

$$\langle \tilde{\varphi}(x_r), \lambda f \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \lambda \sum_{j=1}^{n} a_j \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_r},$$

$$\langle \tilde{\varphi}(x_r), Cf \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \frac{1}{n} \sum_{i=1}^{n} \tilde{\varphi}(x_i) \sum_{j=1}^{n} a_j \tilde{k}(x_i, x_j) \right\rangle_{\mathcal{H}}$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^n \tilde{\varphi}(x_i)\sum_{j=1}^n a_j \tilde{k}(x_i, x_j)$.
- Idea: multiple by $\tilde{\varphi}(x_r)$

$$\langle \tilde{\varphi}(x_r), \lambda f \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \lambda \sum_{j=1}^n a_j \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^n a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_r},$$

$$\langle \tilde{\varphi}(x_r), Cf \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \frac{1}{n}\sum_{i=1}^n \tilde{\varphi}(x_i) \sum_{j=1}^n a_j \tilde{k}(x_i, x_j) \right\rangle_{\mathcal{H}}$$

$$= \frac{1}{n}\sum_{i=1}^n \tilde{G}_{ri} \underbrace{\sum_{j=1}^n a_j \tilde{G}_{ij}}_{(\tilde{\mathbf{G}}\mathbf{a})_i}$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^n \tilde{\varphi}(x_i)\sum_{j=1}^n a_j \tilde{k}(x_i, x_j)$.
- Idea: multiple by $\tilde{\varphi}(x_r)$

$$\langle \tilde{\varphi}(x_r), \lambda f\rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \lambda \sum_{j=1}^n a_j \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^n a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_r},$$

$$\langle \tilde{\varphi}(x_r), Cf\rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \frac{1}{n}\sum_{i=1}^n \tilde{\varphi}(x_i) \sum_{j=1}^n a_j \tilde{k}(x_i, x_j) \right\rangle_{\mathcal{H}}$$

$$= \frac{1}{n}\sum_{i=1}^n \tilde{G}_{ri} \underbrace{\sum_{j=1}^n a_j \tilde{G}_{ij}}_{(\tilde{\mathbf{G}}\mathbf{a})_i} = \frac{1}{n}(\tilde{\mathbf{G}}^2\mathbf{a})_r.$$

Eigenvalue problem

- We want to solve $Cf = \lambda f$, $Cf = \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\sum_{j=1}^{n} a_j \tilde{k}(x_i, x_j)$.
- Idea: multiple by $\tilde{\varphi}(x_r)$

$$\langle \tilde{\varphi}(x_r), \lambda f \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \lambda \sum_{j=1}^{n} a_j \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}} = \lambda \underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{rj}}_{(\tilde{\mathbf{G}}\mathbf{a})_r},$$

$$\langle \tilde{\varphi}(x_r), Cf \rangle_{\mathcal{H}} = \left\langle \tilde{\varphi}(x_r), \frac{1}{n}\sum_{i=1}^{n}\tilde{\varphi}(x_i)\sum_{j=1}^{n} a_j \tilde{k}(x_i, x_j) \right\rangle_{\mathcal{H}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\tilde{G}_{ri}\underbrace{\sum_{j=1}^{n} a_j \tilde{G}_{ij}}_{(\tilde{\mathbf{G}}\mathbf{a})_i} = \frac{1}{n}(\tilde{\mathbf{G}}^2\mathbf{a})_r.$$

- Eigenvalue problem: $\tilde{\mathbf{G}}^2\mathbf{a} = n\lambda\tilde{\mathbf{G}}\mathbf{a}$, i.e. $\tilde{\mathbf{G}}\mathbf{a} = (n\lambda)\mathbf{a}$.

Orthogonal eigenvectors in kernel PCA

Taking two (eigenvector, eigenvalue) pairs:

$$f_1 = \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \qquad \tilde{\mathbf{G}}\mathbf{a}_1 = \lambda_1\mathbf{a}_1,$$

$$f_2 = \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j), \qquad \tilde{\mathbf{G}}\mathbf{a}_2 = \lambda_2\mathbf{a}_2.$$

one has

$$0 \stackrel{?}{=} \langle f_1, f_2 \rangle_{\mathcal{H}}$$

Orthogonal eigenvectors in kernel PCA

Taking two (eigenvector, eigenvalue) pairs:

$$f_1 = \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \qquad \tilde{\mathbf{G}}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1,$$

$$f_2 = \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j), \qquad \tilde{\mathbf{G}}\mathbf{a}_2 = \lambda_2 \mathbf{a}_2.$$

one has

$$0 \overset{?}{=} \langle f_1, f_2 \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}}$$

Orthogonal eigenvectors in kernel PCA

Taking two (eigenvector, eigenvalue) pairs:

$$f_1 = \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \qquad \tilde{\mathbf{G}}\mathbf{a}_1 = \lambda_1\mathbf{a}_1,$$

$$f_2 = \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j), \qquad \tilde{\mathbf{G}}\mathbf{a}_2 = \lambda_2\mathbf{a}_2.$$

one has

$$0 \stackrel{?}{=} \langle f_1, f_2\rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}} = \mathbf{a}_1^T \tilde{\mathbf{G}}\mathbf{a}_2$$

Orthogonal eigenvectors in kernel PCA

Taking two (eigenvector, eigenvalue) pairs:

$$f_1 = \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \qquad \tilde{\mathbf{G}}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1,$$

$$f_2 = \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j), \qquad \tilde{\mathbf{G}}\mathbf{a}_2 = \lambda_2 \mathbf{a}_2.$$

one has

$$0 \stackrel{?}{=} \langle f_1, f_2 \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} a_{1i}\tilde{\varphi}(x_i), \sum_{j=1}^{n} a_{2j}\tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}} = \mathbf{a}_1^T \tilde{\mathbf{G}}\mathbf{a}_2 = \mathbf{a}_1^T \lambda_2 \mathbf{a}_2.$$

Orthogonality $\Rightarrow$ projection is easy

Projection of a new x^* to the first d-PCs:

$$\Pi\left[\tilde{\varphi}\left(x^*\right)\right] = \sum_{j=1}^{d} \left\langle \tilde{\varphi}\left(x^*\right), f_j \right\rangle_{\mathcal{H}} f_j.$$

Orthogonality ⇒ projection is easy

Projection of a new x^* to the first d-PCs:

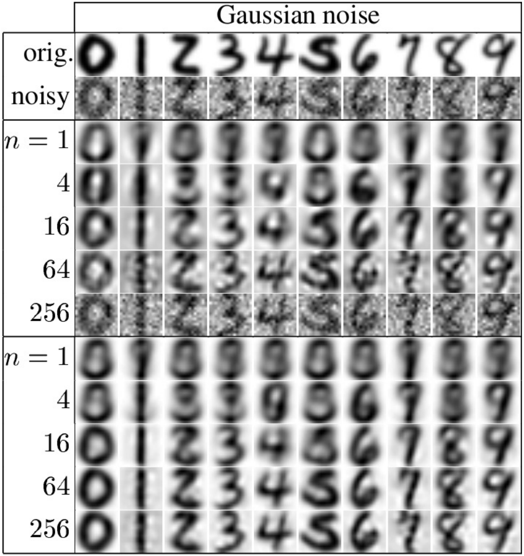
$$\Pi\left[\tilde{\varphi}\left(x^*\right)\right] = \sum_{j=1}^{d} \left\langle \tilde{\varphi}\left(x^*\right), f_j \right\rangle_{\mathcal{H}} f_j.$$

For fixed $f = f_j$, using $f = \sum_{i=1}^{n} a_i \tilde{\varphi}(x_i)$:

$$\left\langle \tilde{\varphi}\left(x^*\right), f \right\rangle_{\mathcal{H}} f = \sum_{i} a_i \tilde{k}\left(x_i, x^*\right) f = \sum_{i,j=1}^{n} a_i a_j \tilde{k}\left(x_i, x^*\right) \tilde{\varphi}(x_j).$$

In denoising application: PCA vs kernel PCA

The pre-image problem to solve: $\widehat{x^*} = \arg\min_{x\in\mathcal{X}} \|\tilde{\varphi}(x) - \Pi[\tilde{\varphi}(x^*)]\|^2_{\mathcal{H}}$.



Kernel-based Divergence & Independence Measures

KL Divergence and Mutual Information Alternatives

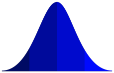
- Mean embedding:

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

KL Divergence and Mutual Information Alternatives

- Mean embedding:

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k = \overline{span}\left(k(\cdot,x) : x \in \mathcal{X}\right).$$



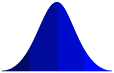
- Maximum mean discrepancy:

$$\text{MMD}_k(\mathbb{P},\mathbb{Q}) = \left\|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\right\|_{\mathcal{H}_k}.$$

KL Divergence and Mutual Information Alternatives

- Mean embedding:

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k = \overline{span}\left(k(\cdot,x) : x \in \mathcal{X}\right).$$



- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion, $k = k_1 \otimes k_2$:

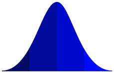
$$\mathrm{HSIC}_k\left(\mathbb{P}\right) = \mathrm{MMD}_k\left(\mathbb{P}, \mathbb{P}_1 \otimes \mathbb{P}_2\right),$$

$$(k_1 \otimes k_2)\left((x,y),(x',y')\right) = k_1(x,x')k_2(y,y').$$

KL Divergence and Mutual Information Alternatives

- Mean embedding:

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot,x)} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k = \overline{span}\left(k(\cdot,x) : x \in \mathcal{X}\right).$$



- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

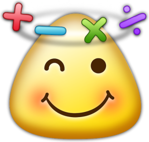
- Hilbert-Schmidt independence criterion, $k = k_1 \otimes k_2$:

$$\mathrm{HSIC}_k(\mathbb{P}) = \mathrm{MMD}_k(\mathbb{P}, \mathbb{P}_1 \otimes \mathbb{P}_2),$$
$$(k_1 \otimes k_2)\left((x,y),(x',y')\right) = k_1(x,x')k_2(y,y').$$

- Kernel Canonical Correlation Analysis:

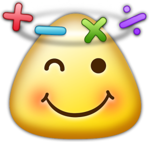
$$\mathrm{KCCA}(\mathbb{P}_{xy}) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \mathrm{corr}(f(x), g(y)).$$

- 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 – History of KCCA

- 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:

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

satisfies 1-4.

Independence measures

- He showed:

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

  satisfies 1-4.
- Too ambitious:
  - computationally intractable.
  - many functions.

Independence measures: restriction to continuous functions

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

Independence measures: restriction to continuous functions

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R}$, bounded continuous$\}$ would also work.
- Still too large!
- Idea:
  - certain $\mathcal{H}_k$ function classes are dense in $C_b(\mathcal{X})$.
  - computationally tractable.

Wanted

- Independence measure,
- distance,
- inner product

measures/estimates on probability distributions

without density estimation!

Kernel Canonical Correlation Analysis (KCCA)

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

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

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

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:

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

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:

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

KCCA: empirical estimate

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

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

Similarly:

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

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

KCCA: empirical estimate

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

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g as $\left\langle g, \tilde{\psi}(y_n)\right\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ 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 as $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ 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 $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

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

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

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

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

and we have

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

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

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

Thus,

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

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

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 $a,b \in \mathbb{R}\backslash\{0\}$.
- 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., 2005].

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., 2005].
- 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., 2005].
- Enough: universal kernel on a compact metric domain (later).
- Example ($\gamma > 0$):
  - Gaussian: $k(x, x') = e^{-\gamma \left\| x - x' \right\|_2^2}$.
  - Laplacian kernel: $k(x, x') = e^{-\gamma \left\| x - x' \right\|_2}$.

KCCA: regularization

In fact, we estimated

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

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

KCCA: regularization

In fact, we estimated

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

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

- 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., 2005], [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.

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

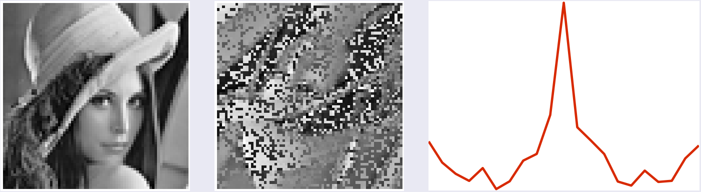
In short

$\tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}$ 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 ($\mathbf{H}^2 = \mathbf{H}$).

Recall: outlier-robust image registration (it was KCCA)



KCCA: finished.

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

Mean embedding: 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).

Mean embedding: 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).

Intuition of MMD and HSIC estimation follows

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k},$$
$$\mathrm{HSIC}_k(\mathbb{P}) = \mathrm{MMD}_k(\mathbb{P}, \mathbb{P}_1 \otimes \mathbb{P}_2).$$

Maximum Mean Discrepancy (MMD)

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

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

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

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

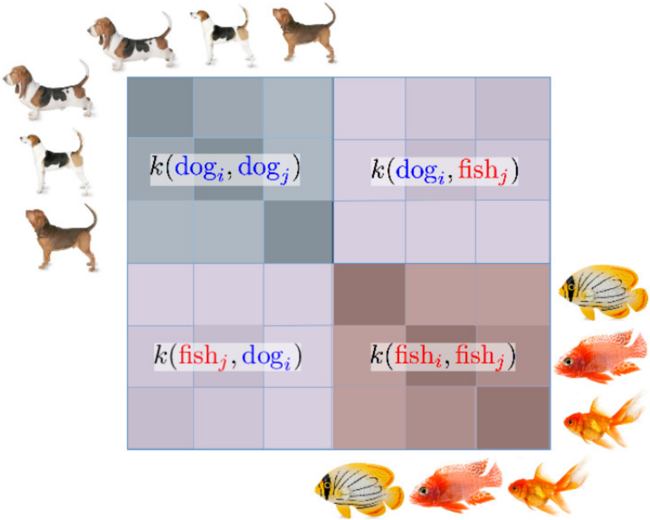
$k(x,y)$	$K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k}$
$e^{-\frac{\gamma}{2}\|x-y\|_2^2}$	$\frac{e^{-\frac{1}{2}(m_1-m_2)^T(\Sigma_1+\Sigma_2+\gamma I)^{-1}(m_1-m_2)}}{\|\gamma\Sigma_1+\gamma\Sigma_2+I\|^{\frac{1}{2}}}$
$(1+\langle x,y\rangle)^2$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

MMD estimator: intuition

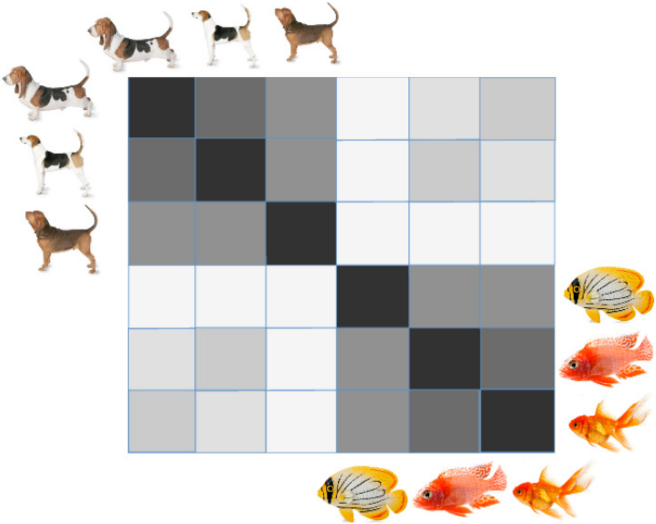
$\sim P$

$\sim Q$

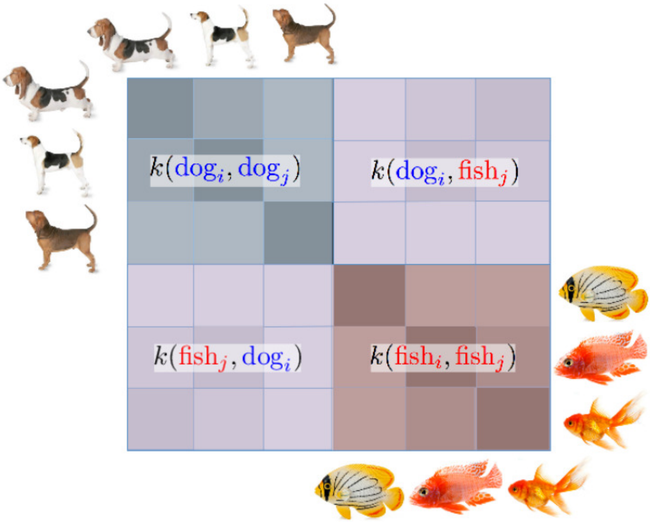
MMD estimator: intuition



MMD estimator: intuition



MMD estimator: intuition



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

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

MMD estimator: mean of kernel values

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}} \varphi(x)$.

MMD estimator: mean of kernel values

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.
- MMD = difference between feature means:

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

MMD estimator: mean of kernel values

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}} \varphi(x)$.
- MMD = difference between feature means:

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

MMD estimator: mean of kernel values

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.
- MMD = difference between feature means:

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

MMD estimator: mean of kernel values

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.
- MMD = difference between feature means:

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

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

using $\{x_i\}_{i=1}^m \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$ samples.

- Computational complexity: $\mathcal{O}\left((m+n)^2\right)$, quadratic.

Hilbert-Schmidt Independence Criterion (HSIC)

HSIC: intuition. $\mathcal{X}$: images, $\mathcal{Y}$: descriptions

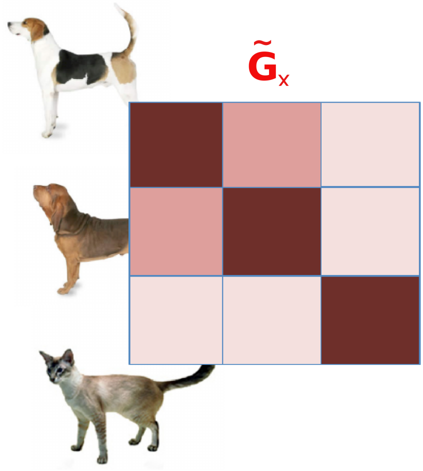
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

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.

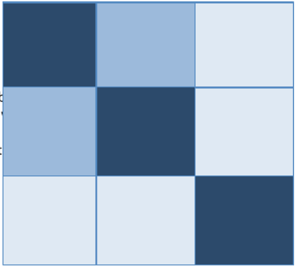
Text from dogtime.com and petfinder.com

HSIC intuition: Gram matrices



$\tilde{G}_x$

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.

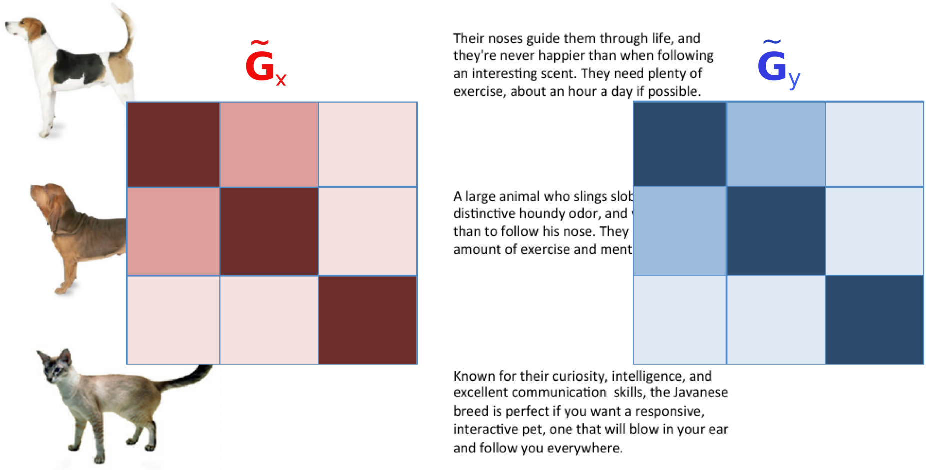


$\tilde{G}_y$

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.

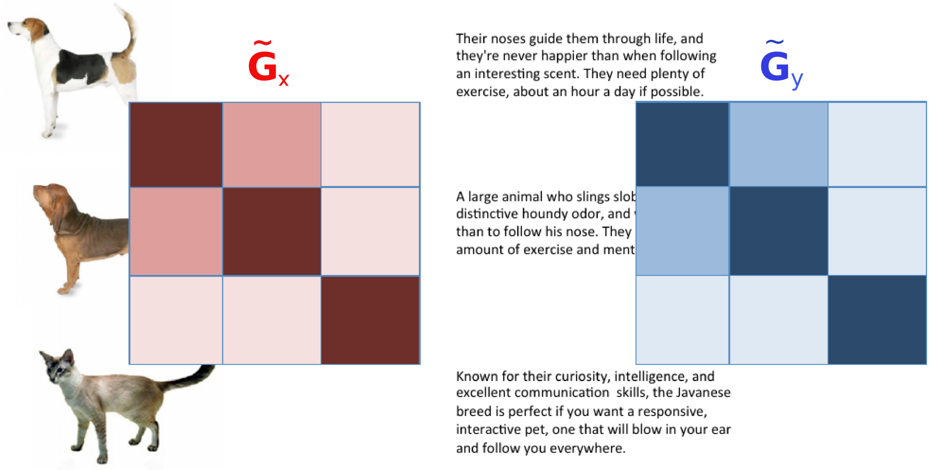
HSIC intuition: Gram matrices



Empirical estimate:

$$\widehat{\mathrm{HSIC}^2} = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F.$$

HSIC intuition: Gram matrices



Empirical estimate:

$$\widehat{\mathrm{HSIC}^2} = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F. \quad \mathrm{HSIC}(\mathbb{P}_{xy}) = \mathrm{MMD}\left(\mathbb{P}_{xy}, \mathbb{P}_x \otimes \mathbb{P}_y\right).$$

Idea of the HSIC estimator

MMD in terms of kernel evaluations:

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}}k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}}k(y,y') - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}k(x,y).$$

Idea of the HSIC estimator

MMD in terms of kernel evaluations:

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}} k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}} k(y,y') - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}} k(x,y).$$

Question

Can we rewrite HSIC in terms of expected kernel values?

Idea of the HSIC estimator

MMD in terms of kernel evaluations:

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}}k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}}k(y,y') - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}k(x,y).$$

Question

Can we rewrite HSIC in terms of expected kernel values?

$$\mathrm{HSIC}^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right].$$

Idea of the HSIC estimator

MMD in terms of kernel evaluations:

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P},x'\sim\mathbb{P}}k(x,x') + \mathbb{E}_{y\sim\mathbb{Q},y'\sim\mathbb{Q}}k(y,y') - 2\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}k(x,y).$$

Question

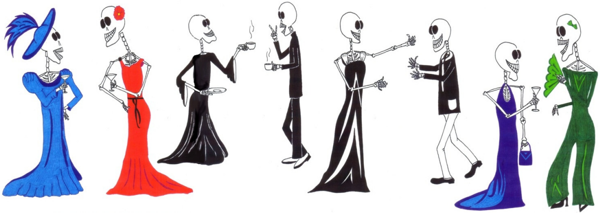
Can we rewrite HSIC in terms of expected kernel values?

$$\mathrm{HSIC}^2(x,y) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}k(x,x')\ell(y,y') + \mathbb{E}_{xx'}k(x,x')\mathbb{E}_{yy'}\ell(y,y') - 2\mathbb{E}_{xy}\left[\mathbb{E}_{x'}k(x,x')\mathbb{E}_{y'}\ell(y,y')\right].$$

Empirical estimation results in

$$\widehat{\mathrm{HSIC}_b^2}(x,y) = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_x \right\rangle_F.$$

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: $\{\mathbf{x}_t\}_{t=1}^T \rightarrow \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T,$

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.

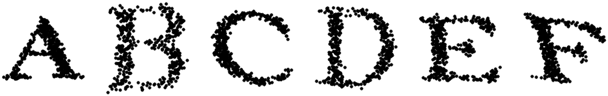
- Goal: $\{\mathbf{x}_t\}_{t=1}^T \to \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,
- Objective function:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x},$$

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

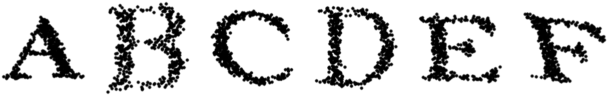
ISA: source, observation

- Hidden sources ($\mathbf{s}$):

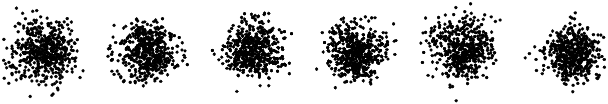


ISA: source, observation

- Hidden sources ($\mathbf{s}$):



- Observation ($\mathbf{x}$):



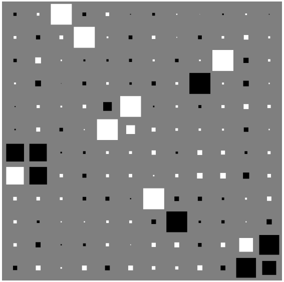
ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

- Performance ($\hat{\mathbf{W}}\mathbf{A}$), ambiguity:

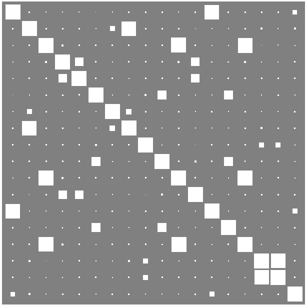


Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation.

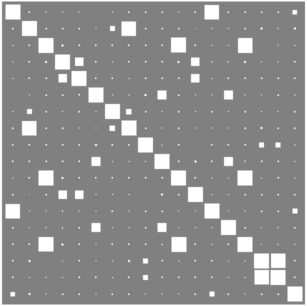
Conjecture: ISA separation theorem [Cardoso, 1998]

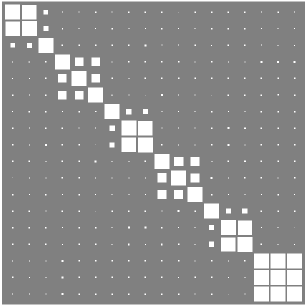
- ISA = ICA + permutation. $\widehat{\text{HSIC}}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.



Conjecture: ISA separation theorem [Cardoso, 1998]

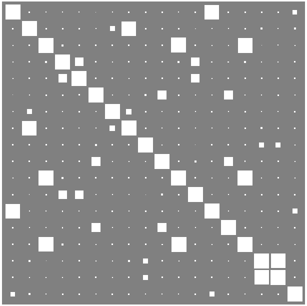
- ISA = ICA + permutation. $\widehat{\text{HSIC}}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.

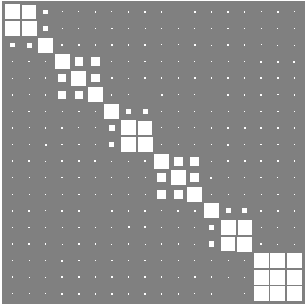




Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation. $\widehat{\text{HSIC}}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.

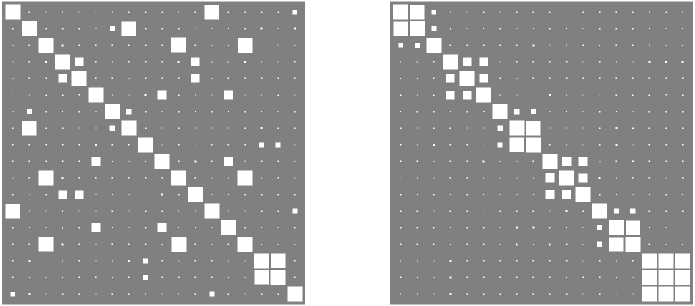




- Basis of the state-of-the-art ISA solvers.

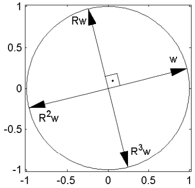
Conjecture: ISA separation theorem [Cardoso, 1998]

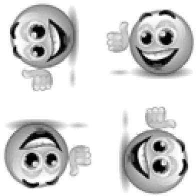
- ISA = ICA + permutation. $\widehat{\text{HSIC}}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.



- Basis of the state-of-the-art ISA solvers.
- Sufficient conditions [Szabó et al., 2012]:
  - $\mathbf{s}^m$: spherical.

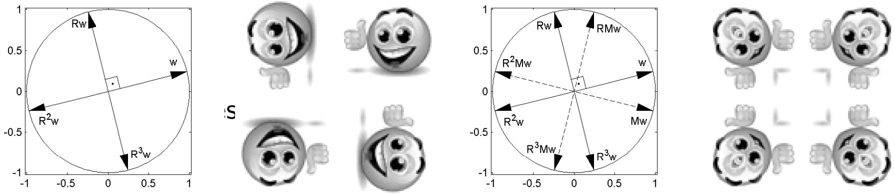
ISA separation theorem





Invariance to

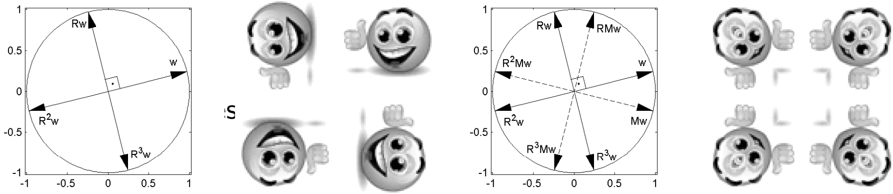
- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.



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: $f(u_1, u_2) = h\left(\sum_i |u_i|^p\right) \quad (p > 0)$.

Universal kernel (see KCCA)

Universality

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 $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Universality

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 $\Rightarrow$

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

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, $\alpha > 0$

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

Universal kernels, $\alpha > 0$

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

Universal kernels, $\alpha > 0$

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

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

Universality: notes

- k: universal $\Leftrightarrow$ $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective on finite signed measures (Hahn-Banach).

Universality: notes

- k: universal $\Leftrightarrow \mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective on finite signed measures (Hahn-Banach).
- Thus, universal $\Rightarrow$ characteristic.

Universality: notes

- k: universal $\Leftrightarrow \mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective on finite signed measures (Hahn-Banach).
- Thus, universal $\Rightarrow$ characteristic.
- 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].

Characteristic property, i.e. when MMD is a metric?

Polynomial kernels are not characteristic

[Sriperumbudur et al., 2010b]:

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

$$\text{MMD}_k^2(\mathbb{P},\mathbb{Q}) = \left\|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\right\|^2, \qquad \mathbf{m}_{\mathbb{P}} = \int_{\mathcal{X}} \mathbf{x} \mathrm{d}\mathbb{P}(x).$$

Polynomial kernels are not characteristic

[Sriperumbudur et al., 2010b]:

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

$$\text{MMD}_k^2(\mathbb{P},\mathbb{Q}) = \left\|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\right\|^2, \qquad \mathbf{m}_{\mathbb{P}} = \int_{\mathcal{X}} \mathbf{x}\mathrm{d}\mathbb{P}(x).$$

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

$$\text{MMD}_k^2(\mathbb{P},\mathbb{Q}) = 2\left\|\mathbf{m}_{\mathbb{P}} - \mathbf{m}_{\mathbb{Q}}\right\|^2 + \left\|\boldsymbol{\Sigma}_{\mathbb{P}} - \boldsymbol{\Sigma}_{\mathbb{Q}} + \mathbf{m}_{\mathbb{P}}\mathbf{m}_{\mathbb{P}}^T - \mathbf{m}_{\mathbb{Q}}\mathbf{m}_{\mathbb{Q}}^T\right\|_F^2,$$

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

MMD in terms of characteristic functions

Using Bochner's theorem:

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

MMD in terms of characteristic functions

Using Bochner's theorem:

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

MMD in terms of characteristic functions

Using Bochner's theorem:

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

$$= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle}\mathrm{d}\Lambda(\boldsymbol{\omega})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})$$

$$= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x},\boldsymbol{\omega}\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\right]}_{\overline{c_{\mathbb{P}}(\boldsymbol{\omega})-c_{\mathbb{Q}}(\boldsymbol{\omega})}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle \mathbf{y},\boldsymbol{\omega}\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\right]}_{c_{\mathbb{P}}(\boldsymbol{\omega})-c_{\mathbb{Q}}(\boldsymbol{\omega})} \mathrm{d}\Lambda(\boldsymbol{\omega})$$

MMD in terms of characteristic functions

Using Bochner's theorem:

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

MMD in terms of characteristic functions

Using Bochner's theorem:

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

Simple description for shift-invariant kernels on $\mathbb{R}^d$

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

k is characteristic iff. supp$(\Lambda) = \mathbb{R}^d$, where

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

Simple description for shift-invariant kernels on $\mathbb{R}^d$

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

k is characteristic iff. $supp(\Lambda) = \mathbb{R}^d$, where

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

Example on $\mathbb{R}$:

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma, \sigma]$

Simple description for shift-invariant kernels on $\mathbb{R}^d$

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

k is characteristic iff. $supp(\Lambda) = \mathbb{R}^d$, where

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

Example on $\mathbb{R}$:

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma, \sigma]$

Similar characterization $\exists$ on 'Bochner domains' [Berg et al., 1984, Fukumizu et al., 2009].

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 $\mathcal{H}_k$.

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

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 $\mathcal{H}_k$.

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

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 $\mathcal{H}_k$.

$$\begin{aligned}
\text{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} \\
&= \sup_{f \in \mathcal{F}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} \\
&= \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{Q}f), \quad \mathbb{P}f := \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{P}(x).
\end{aligned}$$

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 $\mathcal{H}_k$.

$$\begin{aligned} \text{MMD}(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} \\ &= \sup_{f \in \mathcal{F}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_k} \\ &= \sup_{f \in \mathcal{F}} (\mathbb{P}f - \mathbb{Q}f), \quad \mathbb{P}f := \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{P}(x). \end{aligned}$$

- 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]:

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

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: easier.

$\mathcal{I}$-characteristic property, i.e. when HSIC is an independence measure?

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 $\overset{\text{[Sriperumbudur et al., 2011]}}{\underset{\text{[Sriperumbudur et al., 2011]}}{\overset{\not\Longrightarrow}{\Longleftarrow}}}$ -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]:

$$k_1 \& k_2\text{: universal} \Rightarrow k_1 \otimes k_2\text{: universal } (\Rightarrow \mathcal{I}\text{-characteristic}).$$

- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013b]:

$$k_1 \& k_2\text{: characteristic} \Leftrightarrow k_1 \otimes k_2\text{: } \mathcal{I}\text{-characteristic}.$$

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]:
  $$k_1 \& k_2\text{: characteristic} \Leftrightarrow k_1 \otimes k_2\text{: } \mathcal{I}\text{-characteristic}.$$

Question

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]:

$$k_1 \& k_2\text{: characteristic} \Leftrightarrow k_1 \otimes k_2\text{: } \mathcal{I}\text{-characteristic}.$$

Question

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, 2018]

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, 2018]

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, 2018]

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, 2018]

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)

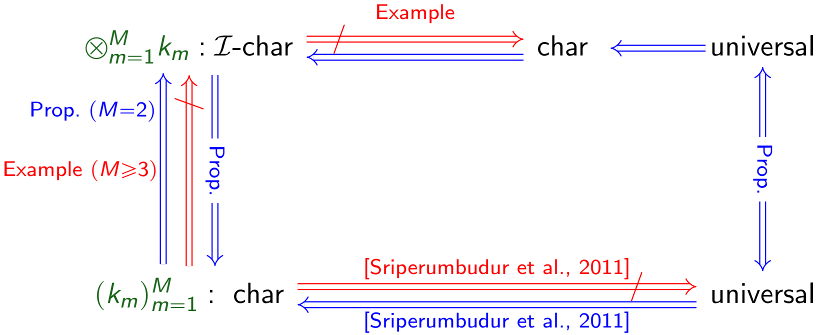
- k_1, k_2: *characteristic* $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$*-characteristic.*
- $\Leftarrow$: *for* $\forall M \geqslant 2$.
- k_1, k_2, k_3: *characteristic* $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$*-characteristic [Ex].*

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.*

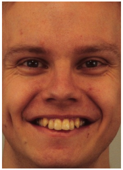


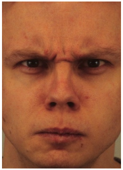
$\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-char
Example
char
universal
Prop. (M=2)
Example ($M \geqslant 3$)
Prop.
Prop.
$(k_m)_{m=1}^{M}$: char
[Sriperumbudur et al., 2011]
universal
[Sriperumbudur et al., 2011]

Hypothesis Testing

Two-sample testing: recall

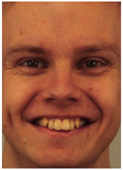
- Given:
  - $X = \{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \sim \mathbb{Q}$.
  - Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.

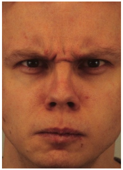




Two-sample testing: recall

- Given:
  - $X = \{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \sim \mathbb{Q}$.
  - Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.



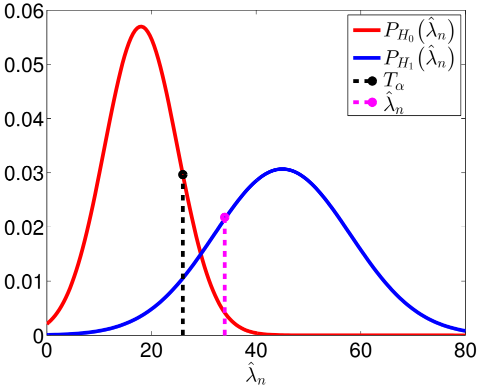


- Problem: using X, Y test

$$H_0 : \mathbb{P} = \mathbb{Q}, \text{ vs}$$
$$H_1 : \mathbb{P} \neq \mathbb{Q}.$$

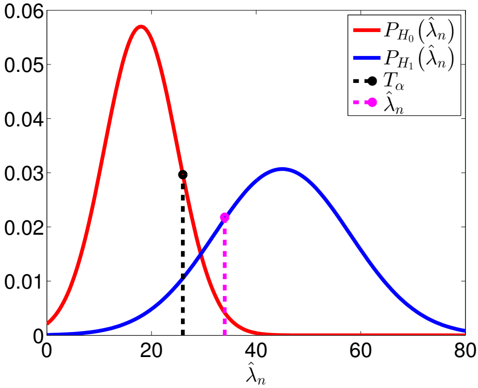
Ingredients of two-sample test

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- Under H_0: $P_{H_0}(\underbrace{\hat{\lambda}_n \leqslant T_\alpha}_{\text{correctly accepting } H_0}) = 1 - \alpha$.



Ingredients of two-sample test

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- Under H_0: $P_{H_0}(\underbrace{\hat{\lambda}_n \leqslant T_\alpha}_{\text{correctly accepting } H_0}) = 1 - \alpha$.
- Under H_1: $P_{H_1}(T_\alpha < \hat{\lambda}_n) = P(\text{correctly rejecting } H_0) =: \text{power}$.



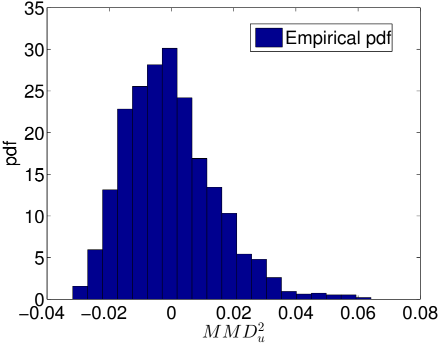
Two-sample test using MMD asymptotics: H_0

Under H_0 [Gretton et al., 2007, Gretton et al., 2012] $\xrightarrow{\text{U-statistics}}$

$$n\widehat{\mathrm{MMD}}_u^2(\mathbb{P},\mathbb{P}) \sim \sum_{i=1}^{\infty} \lambda_i \left(z_i^2 - 2\right),$$

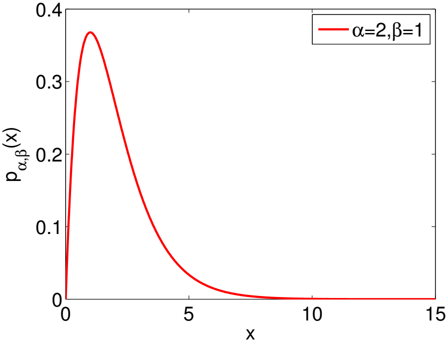
where $z_i \sim N(0,2)$ i.i.d.,

$$\int_{\mathcal{X}} \tilde{k}(x,x')v_i(x)\mathrm{d}\mathbb{P}(x) = \lambda_i v_i(x'),\ \tilde{k}(x,x') = \left\langle \varphi(x) - \mu_{\mathbb{P}}, \varphi(x') - \mu_{\mathbb{P}} \right\rangle_{\mathcal{H}_k}.$$



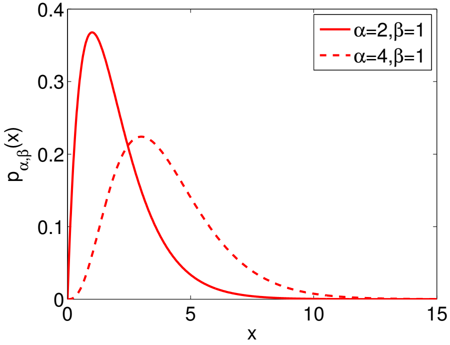
Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



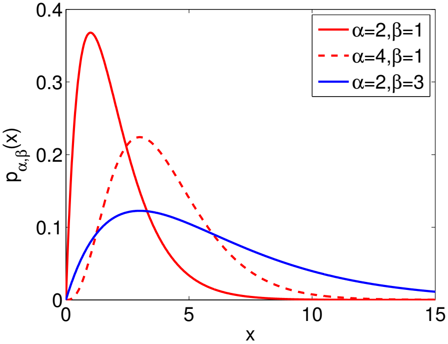
Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



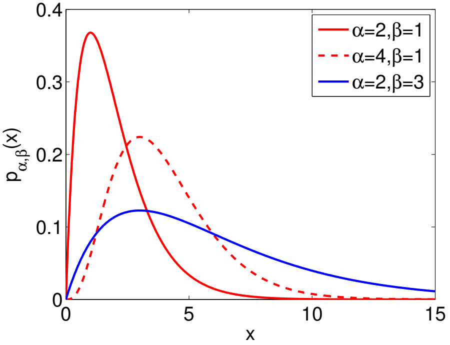
Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
    - gamma approximation:



Null approximations; test statistics: quadratic in time

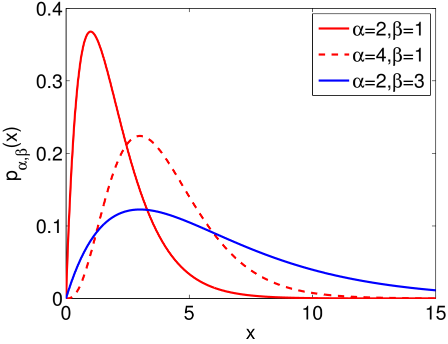
- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



  - truncated expansion [Gretton et al., 2009].

Null approximations; test statistics: quadratic in time

- 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] (large var),
  - recent linear methods (soon).

Independence testing with HSIC

Similary story [Gretton et al., 2008, Pfister et al., 2017]:

- Null asymptotics:

$$\sum_{i=1}^{\infty} \lambda_i z_i^2, \quad z_i \sim N(0,1).$$

- In practice: permutation-test/gamma-approximation.

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
    - var $\searrow$

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - var $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013a].

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - var $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013a].
- Goodness-of-fit [Chwialkowski et al., 2016].

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - var $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013a].
- Goodness-of-fit [Chwialkowski et al., 2016].
- Time-series:
  - independence (stationary $\rightarrow$ shift) [Chwialkowski and Gretton, 2014],
  - wild bootstrap: [Chwialkowski et al., 2014, Rubenstein et al., 2016].

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - var $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013a].
- Goodness-of-fit [Chwialkowski et al., 2016].
- Time-series:
  - independence (stationary $\rightarrow$ shift) [Chwialkowski and Gretton, 2014],
  - wild bootstrap: [Chwialkowski et al., 2014, Rubenstein et al., 2016].
- block-HSIC [Zhang et al., 2017]:
  - RFF acceleration.

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - var $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013a].
- Goodness-of-fit [Chwialkowski et al., 2016].
- Time-series:
  - independence (stationary $\rightarrow$ shift) [Chwialkowski and Gretton, 2014],
  - wild bootstrap: [Chwialkowski et al., 2014, Rubenstein et al., 2016].
- block-HSIC [Zhang et al., 2017]:
  - RFF acceleration.
- Conditional independence & RFF [Strobl et al., 2017].

Linear-time Tests

Linear-time 'MMD'

Idea [Chwialkowski et al., 2015]

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

$$\rho(\mathbb{P}, \mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}, \qquad \mathcal{V} = \{\mathbf{v}_j\}_{j=1}^{J},$$

Linear-time 'MMD'

Idea [Chwialkowski et al., 2015]

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

Plug-in estimate: $\mathcal{O}(n)$-time.

$$\rho(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}, \qquad \mathcal{V} = \{\mathbf{v}_j\}_{j=1}^{J},$$

$$\hat{\rho}(\mathbb{P},\mathbb{Q}) = \frac{\bar{\mathbf{z}}_n^T \bar{\mathbf{z}}_n}{J}, \qquad \bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n} \underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^{J}}_{=:\mathbf{z}(\mathbf{x}_i,\mathbf{y}_i)},$$

Linear-time 'MMD'

Idea [Chwialkowski et al., 2015]

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

Plug-in estimate: $\mathcal{O}(n)$-time. Whitened test statistic: χ^2_J null.

$$\rho(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}, \qquad \mathcal{V} = \{\mathbf{v}_j\}_{j=1}^{J},$$

$$\hat{\rho}(\mathbb{P},\mathbb{Q}) = \frac{\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n}{J}, \qquad \bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}\underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^{J}}_{=:\mathbf{z}(\mathbf{x}_i,\mathbf{y}_i)},$$

$$\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T\mathbf{\Sigma}_n^{-1}\bar{\mathbf{z}}_n, \qquad \mathbf{\Sigma}_n = \widehat{\text{cov}}\left(\{\mathbf{z}(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{n}\right),$$

Linear-time 'MMD'

Idea [Chwialkowski et al., 2015], [Jitkrittum et al., 2016]

Replace $\|\cdot\|_{\mathcal{H}_k}$ in MMD with $\|\cdot\|_{L^2(\mathcal{V})}$. Metric a.s. for analytic & characteristic $k = k_\sigma$.

Plug-in estimate: $\mathcal{O}(n)$-time. Whitened test statistic: χ^2_J null. Power opt.

$$\rho(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}, \qquad \mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J,$$

$$\hat{\rho}(\mathbb{P},\mathbb{Q}) = \frac{\bar{\mathbf{z}}_n^T \bar{\mathbf{z}}_n}{J}, \qquad \bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}\underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^J}_{=:\mathbf{z}(\mathbf{x}_i,\mathbf{y}_i)},$$

$$\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T \mathbf{\Sigma}_n^{-1}\bar{\mathbf{z}}_n, \qquad \mathbf{\Sigma}_n = \widehat{\text{cov}}\left(\{\mathbf{z}(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n\right),$$

$$(\sigma^*,\mathcal{V}^*) = \arg\max_{\sigma,\mathcal{V}} \lambda, \qquad \lambda = n\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m}.$$

Linear-time 'HSIC' [Jitkrittum et al., 2017a]

Use different norm of the witness function (u):

$$\text{HSIC}(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

Linear-time 'HSIC' [Jitkrittum et al., 2017a]

Use different norm of the witness function (u):

$$\text{HSIC}(x,y) = \left\| \mu_{xy} - \mu_x \otimes \mu_y \right\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$\text{FSIC}(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

Linear-time 'HSIC' [Jitkrittum et al., 2017a]

Use different norm of the witness function (u):

$$\text{HSIC}(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$\text{FSIC}(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \left\{(\mathbf{v}_j,\mathbf{w}_j)\right\}_{j=1}^{J},$$

$$= \|u\|_{L^2(\mathcal{V})}.$$

Linear-time 'HSIC' [Jitkrittum et al., 2017a]

Use different norm of the witness function (u):

$$\mathrm{HSIC}(x,y) = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$\mathrm{FSIC}(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

$$= \|u\|_{L^2(\mathcal{V})}.$$

- Whitening $\Rightarrow \chi^2_J$ null. Computation: $\mathcal{O}(n)$. Power optimization.

Linear-time 'HSIC' [Jitkrittum et al., 2017a]

Use different norm of the witness function (u):

$$\text{HSIC}(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_{k_1 \otimes k_2}}, \quad u(\mathbf{v},\mathbf{w}) = \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}),$$

$$\text{FSIC}(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(\mathbf{v}_j,\mathbf{w}_j)}, \qquad \mathcal{V} = \{(\mathbf{v}_j,\mathbf{w}_j)\}_{j=1}^{J},$$

$$= \|u\|_{L^2(\mathcal{V})}.$$

- Whitening $\Rightarrow \chi^2_J$ null. Computation: $\mathcal{O}(n)$. Power optimization.
- Alternative view: $u(\mathbf{v},\mathbf{w}) = \text{cov}_{\mathbf{xy}}(k_1(\mathbf{x},\mathbf{v}), k_2(\mathbf{y},\mathbf{w})) = (\mathbf{v},\mathbf{w})^{th}$ entry of

$$C_{xy} = \mathbb{E}_{xy}\left[\varphi_1(x) \otimes \varphi_2(y)\right] - \mu_x \otimes \mu_y.$$

We

- assumed analytic, characteristic, bounded kernels.
- replaced the RKHS norm with $L^2(\mathcal{V})$ norm.

In linear-time 'MMD' and 'HSIC', respectively:

$$\mathbb{P} = \mathbb{Q} \Leftrightarrow \quad \mu_{\mathbb{P}-\mathbb{Q}} = 0,$$
$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \Leftrightarrow \quad \mu_{\mathbb{P}-\mathbb{P}_1 \otimes \mathbb{P}_2} = 0.$$

Goodness-of-fit

Let $d = 1$. Stein operator of model p

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Goodness-of-fit

Let $d = 1$. Stein operator of model p

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Goodness-of-fit

Let $d = 1$. Stein operator of model p

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Let us take the unit ball of $\mathcal{H}_k$:

$$\sup_{\|f\|_{\mathcal{H}_k}\leqslant 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \|g\|_{\mathcal{H}_k}, \quad g(v) = \mathbb{E}_{x\sim q}\frac{\partial_x[p(x)k(x,v)]}{p(x)}.$$

Goodness-of-fit [Chwialkowski et al., 2016, Liu et al., 2016]

Let $d = 1$. Stein operator of model p

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Let us take the unit ball of $\mathcal{H}_k$:

$$\sup_{\|f\|_{\mathcal{H}_k}\leqslant 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \|g\|_{\mathcal{H}_k}, \quad g(v) = \mathbb{E}_{x\sim q}\frac{\partial_x[p(x)k(x,v)]}{p(x)}.$$

For universal k:

$$\boxed{p = q \Leftrightarrow g = 0 \text{ (witness)}}.$$

Goodness-of-fit [Jitkrittum et al., 2017a], [Chwialkowski et al., 2016, Liu et al., 2016]

Let $d = 1$. Stein operator of model p

$$(T_p f)(x) = \frac{[p(x)f(x)]'}{p(x)} = [\log p(x)]' f(x) + f'(x).$$

Under $\lim_{|x|\to\infty} f(x)p(x) = 0$ (integration by parts):

$$p = q \Rightarrow \mathbb{E}_{x\sim q}(T_p f)(x) = 0.$$

Let us take the unit ball of $\mathcal{H}_k$:

$$\sup_{\|f\|_{\mathcal{H}_k}\leqslant 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \|g\|_{\mathcal{H}_k}, \quad g(v) = \mathbb{E}_{x\sim q}\frac{\partial_x[p(x)k(x,v)]}{p(x)}.$$

For universal k:

$$\boxed{p = q \Leftrightarrow g = 0 \text{ (witness)}}.$$

$L^2(\mathcal{V})$ trick goes through.

Numerical Illustrations

2-sample testing: parameter settings

- Gaussian kernel (σ). $\alpha = 0.01$. $J = 1$. Repeat 500 trials.
- Report rejection rate of H_0
- Compare 4 methods
  - **ME-full**: Optimize $\mathcal{V}$ and 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

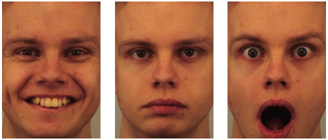
- Aggregating over trials; example: 'Bayes-Neuro'.

- Least discriminative ones:

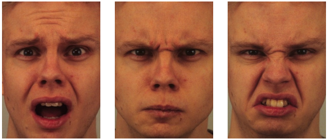
  circumfer, bra, dominiqu, rhino, mitra, kid, impostor.

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.

$+$: 

happy neutral surprised

$-$: 

afraid angry disgusted

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) = 

Independence testing: parameters

- k_1, k_2: 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

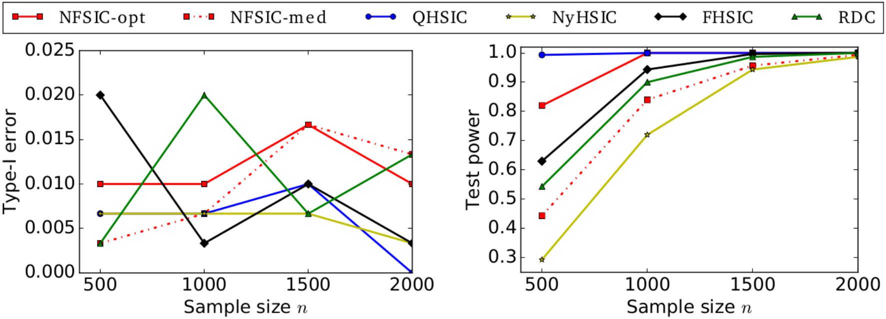
Song (x) vs. year of release (y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $x \in \mathbb{R}^{90}$: audio features.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.

Demo-1: million song data

Song (x) vs. year of release (y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $x \in \mathbb{R}^{90}$: audio features.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.



Demo-2: videos and captions

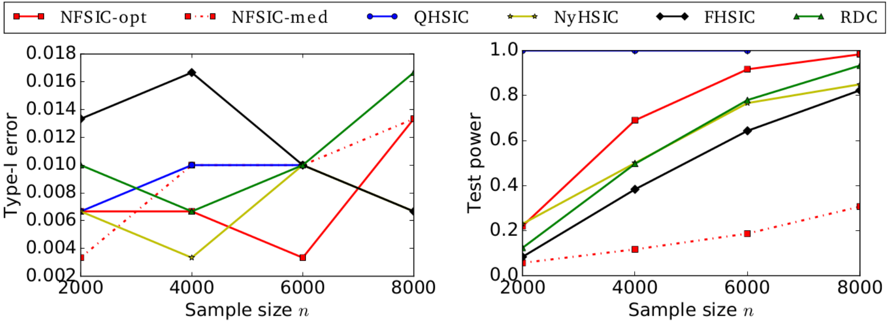
Youtube video (x) vs. caption (y).

- VideoStory46K [Habibian et al., 2014]
- $x \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $y \in \mathbb{R}^{1878}$: bag of words. TF.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.

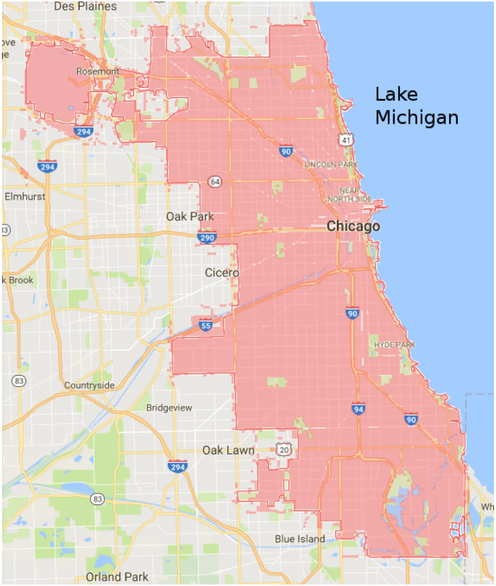
Demo-2: videos and captions

Youtube video (x) vs. caption (y).

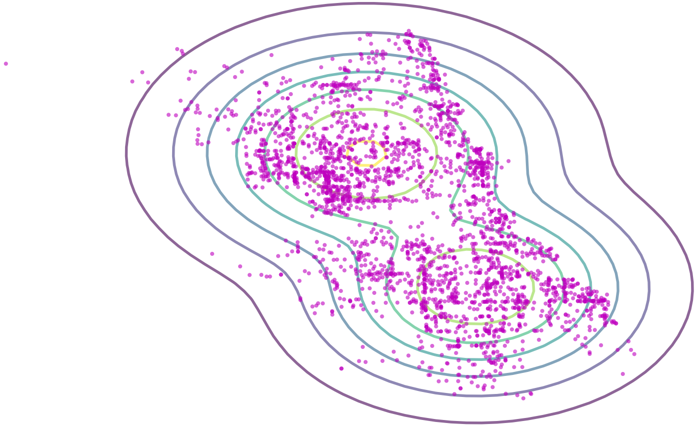
- VideoStory46K [Habibian et al., 2014]
- $x \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $y \in \mathbb{R}^{1878}$: bag of words. TF.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.



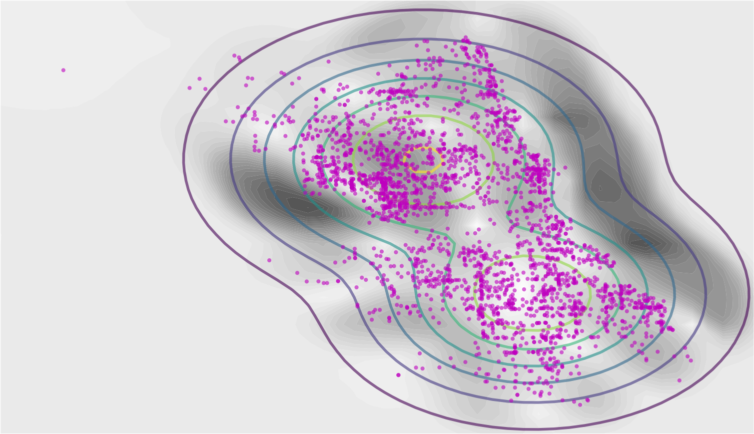
Goodness-of-Fit Demo


Des Plaines
Rosemont
Lake Michigan
294
41
90
LINCOLN PARK
64
NEAR NORTH SIDE
Elmhurst
Oak Park
Chicago
290
Cicero
Brook
90
55
HYDE PARK
83
Countryside
Bridgeview
94
90
Oak Lawn
20
294
83
Blue Island
Orland Park

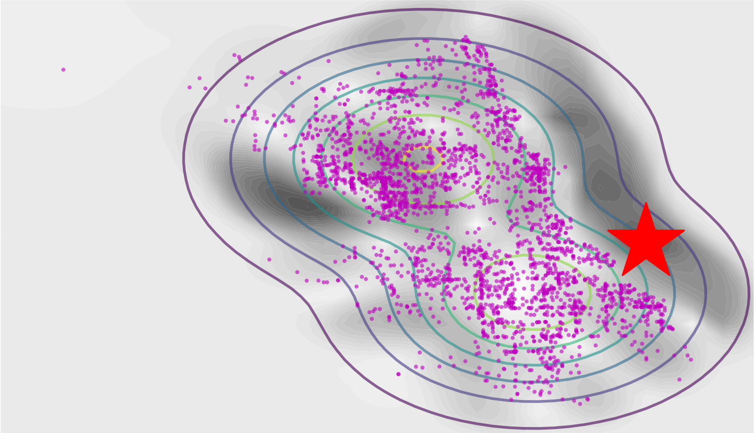
Robbery events (lat/long coordinates) $\sim q$.



Model p: 2-component Gaussian mixture.

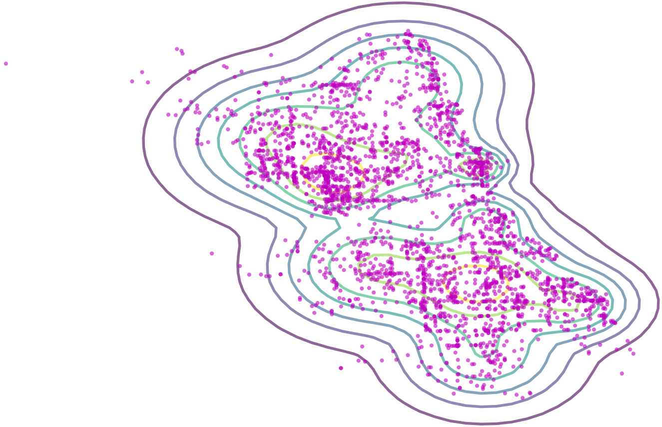


Score surface

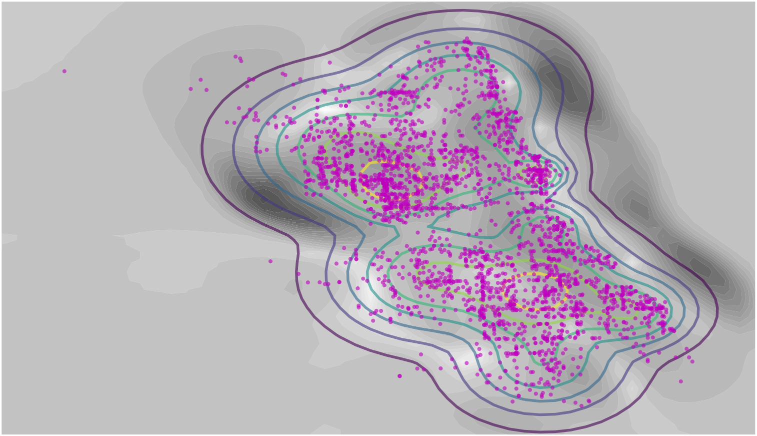


★ = optimized $\mathbf{v}$.

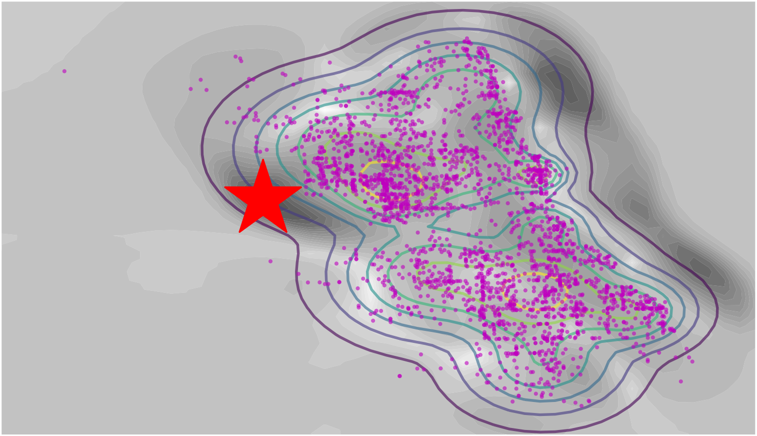
No robbery in Lake Michigan.



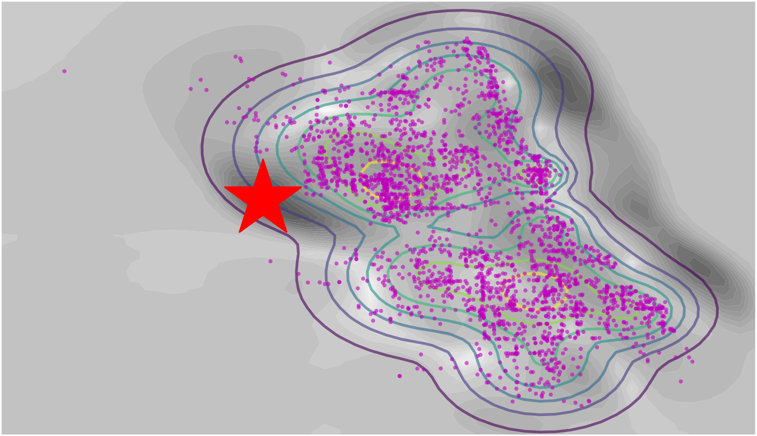
Model p: 10-component Gaussian mixture.



Capture the right tail better.



Still, does not capture the left tail.



Still, does not capture the left tail.

Sharp boundary (geography of Chicago) $\neq$ Gaussian tails. $\rightarrow$ interpretable features

Summary

- Motivation: infoT objectives, hypothesis testing.
- Kernels, RKHS: definitions, construction.
- Kernel applications: classification, ridge regression, PCA.
- MMD, HSIC, KCCA.
- Characteristic, universal, $\mathcal{I}$-characteristic property.
- Hypothesis testing: quadratic & linear-time methods.

Thank you for the attention!

Bach, F. R. and Jordan, M. I. (2002).
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48.

Balasubramanian, K., Li, T., and Yuan, M. (2017).
On the optimality of kernel-embedding based goodness-of-fit tests.
Technical report.
(https://arxiv.org/abs/1709.08148).

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.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011).

The million song dataset.
In *International Conference on Music Information Retrieval (ISMIR)*.

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018).
Demystifying MMD GANs.
In *International Conference on Learning Representations (ICLR)*.

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2017).
Domain generalization by marginal transfer learning.
Technical report.
(https://arxiv.org/abs/1711.07910).

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.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. (2006).
Integrating structured biological data by kernel maximum mean discrepancy.
Bioinformatics, 22(14):e49–e57.

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.

Chwialkowski, K. and Gretton, A. (2014).
A kernel independence test for random processes.

In *International Conference on Machine Learning (ICML; PMLR)*, volume 32, pages 1422–1430.

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.

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.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015).
Training generative neural networks via maximum mean discrepancy optimization.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267.

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. (2009).
Characteristic kernels on groups and semigroups.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 473–480.

Fukumizu, K., Song, L., and Gretton, A. (2013).
Kernel Bayes' rule: Bayesian inference with positive definite kernels.
Journal of Machine Learning Research, 14:3753–3783.

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., 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 *Neural Information Processing Systems (NIPS)*, pages 585–592.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005).
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.

Harchaoui, Z., Bach, F., and Moulines, E. (2007).
Testing for homogeneity with kernel Fisher discriminant analysis.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616.

Harchaoui, Z. and Cappé, O. (2007).
Retrospective mutiple change-point estimation with kernels.
In *IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772.

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 AI 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 (PMLR)*, volume 37, pages 2982–2990.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016).

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. (2017a).
An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML; PMLR)*, volume 70, pages 1742–1751.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017b).
A linear-time kernel goodness-of-fit test.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 261–270.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.

In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kim, B., Khanna, R., and Koyejo, O. O. (2016).
Examples are not enough, learn to criticize! criticism for interpretability.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288.

Klus, S., Bittracher, A., Schuster, I., and Schütte, C. (2019).
A kernel-based approach to molecular conformation analysis.
The Journal of Chemical Physics, 149:244109.

Klus, S., Schuster, I., and Muandet, K. (2018).
Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.
Technical report.
(https://arxiv.org/abs/1712.01572).

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016).
Persistence weighted Gaussian kernel for topological data analysis.
In *International Conference on Machine Learning (ICML)*, pages 2004–2013.

Kybic, J. (2004).
High-dimensional mutual information estimation for image registration.
In *IEEE International Conference on Image Processing (ICIP)*, pages 1779–1782.

Law, H. C. L., Sutherland, D. J., Sejdinovic, D., and Flaxman, S. (2018).
Bayesian approaches to distribution regression.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 1167–1176.

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.

Li, Y., Swersky, K., and Zemel, R. (2015).
Generative moment matching networks.
In *International Conference on Machine Learning (ICML; PMLR)*, pages 1718–1727.

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.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014).
Automatic construction and natural-language description of nonparametric regression models.

In *AAAI Conference on Artificial Intelligence*, pages 1242–1250.

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., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.
International Conference on Machine Learning (ICML; PMLR), 37:1452–1461.

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.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).
Distinguishing cause from effect using observational data: Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2011).
Learning from distributions via support measure machines.
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.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016).
K2-ABC: Approximate Bayesian computation with kernel embeddings.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 51, pages 398–407.

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; PMLR)*, volume 31, pages 507–515.

Raj, A., Law, H. C. L., Sejdinovic, D., and Park, M. (2019).
A differentially private kernel two-sample test.

In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.

Rényi, A. (1959).
On measures of dependence.
Acta Mathematica Academiae Scientiarum Hungaricae, 10:441–451.

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.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).
A generalized representer theorem.
In *Conference on Learning Theory (COLT)*, pages 416–426.

Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., and Peters, J. (2015).
Computing functions of random variables via reproducing kernel Hilbert space representations.
Statistics and Computing, 25(4):755–766.

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.

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).

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. (2011).
Kernel belief propagation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715.

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; PMLR)*, volume 9, pages 781–788.

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.

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.

Strobl, E. V., Visweswaran, S., and Zhang, K. (2017).
Approximate kernel-based conditional independence tests for fast non-parametric causal discovery.
Technical report.
(https://arxiv.org/abs/1702.03877).

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. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18:1–29.

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.

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.

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2018).
Post selection inference with kernels.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 84, pages 152–160.

Yu, Y., Cheng, H., Schuurmans, D., and Szepesvári, C. (2013).
Characterizing the representer theorem.
In *International Conference on Machine Learning (ICML; PMLR)*, volume 28, pages 570–578.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017).
Deep sets.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 3394–3404.

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, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013).
Domain adaptation under target and conditional shift.
Journal of Machine Learning Research, 28(3):819–827.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2017).
Large-scale kernel methods for independence testing.
Statistics and Computing, pages 1–18.

Zolotarev, V. M. (1983).
Probability metrics.
Theory of Probability and its Applications, 28:278–302.