Kernel methods for comparing distributions and detecting dependence

Columbia, 2014

Arthur Gretton

Gatsby Unit, CSML, UCL
Some motivating questions...
Detecting differences in amplitude modulated signals

Samples from P

Samples from Q
Case of discrete domains

- How do you compare distributions...
- ...in a discrete domain? [Read and Cressie, 1988]
Case of discrete domains

- How do you compare distributions... 
- ...in a discrete domain? [Read and Cressie, 1988]

\(X_1:\) Now disturbing reports out of Newfoundland show that the fragile snow crab industry is in serious decline. First the west coast salmon, the east coast salmon and the cod, and now the snow crabs off Newfoundland.

\(X_2:\) To my pleasant surprise he responded that he had personally visited those wharves and that he had already announced money to fix them. What wharves did the minister visit in my riding and how much additional funding is he going to provide for Delaps Cove, Hampton, Port Lorne, ... 

\(Y_1:\) Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

\(Y_2:\) On the grain transportation system we have had the Estey report and the Kroeger report. We could go on and on. Recently programs have been announced over and over by the government such as money for the disaster in agriculture on the prairies and across Canada.

\(P_X = P_Y\) 

... 

Are the pink extracts from the same distribution as the gray ones?
Detecting statistical dependence, continuous domain

- How do you detect dependence...
- ...in a continuous domain?
Detecting statistical dependence, continuous domain

- How do you detect dependence...
- ...in a **continuous** domain?

![Sample from $P_{XY}$](image1)

![Discretized empirical $P_{XY}$](image2)

![Discretized empirical $P_X P_Y$](image3)
Detecting statistical dependence, continuous domain

- How do you detect dependence in a continuous domain?

...in a continuous domain?
Detecting statistical dependence, continuous domain

- How do you detect dependence...
- ...in a continuous domain?

**Problem:** fails even in “low” dimensions! [NIPS07a, ALT08]
  - $X$ and $Y$ in $\mathbb{R}^4$, statistic=Power divergence, samples= 1024, cases where dependence detected=0/500

- Too few points per bin
Detecting statistical dependence, discrete domain

- How do you detect dependence...?
- ...in a discrete domain? [Read and Cressie, 1988]

\[ P_{XY} = P_X P_Y \]

\( X_1: \) Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

\( X_2: \) No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

\( Y_1: \) Honorable sénateurs, ma question s’adresse au leader du gouvernement au Sénat et concerne l’aide financière qu’on a annoncée pour les agriculteurs. La plupart des agriculteurs n’ont encore rien reçu de cet argent.

\( Y_2: \) Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n’a pas réduit le financement qu’il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

Are the French text extracts translations of the English ones?
Outline

• An RKHS metric on the space of probability measures
  – Distance between means in space of features (RKHS)
  – Nonparametric two-sample test...
  – ...for (almost!) any data type: strings, images, graphs, groups (rotation matrices), semigroups (histograms),...

• Dependence detection
  – Covariance and Correlation in feature space

• Relation with energy distance and distance covariance

• Interactions with three (or more) variables
Feature mean difference

- Simple example: 2 Gaussians with different means
- Answer: t-test
Feature mean difference

- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$
Feature mean difference

- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi_x = x^2$
Feature mean difference

- Gaussian and Laplace distributions
- Same mean \textit{and} same variance
- Difference in means using higher order features
• Are $P$ and $Q$ different?
Are \( P \) and \( Q \) different?
Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for $P$ vs $Q$

$$
\text{MMD}(P, Q; F) := \sup_{f \in F} \left[ \mathbb{E}_P f(x) - \mathbb{E}_Q f(y) \right].
$$
- **Maximum mean discrepancy**: smooth function for $\mathbf{P}$ vs $\mathbf{Q}$

\[
\text{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[ \mathbb{E}_P f(x) - \mathbb{E}_Q f(y) \right].
\]
**Function Showing Difference in Distributions**

- **Maximum mean discrepancy**: smooth function for \( P \) vs \( Q \)

\[
\text{MMD}(P, Q; F) := \sup_{f \in F} \left[ E_P f(x) - E_Q f(y) \right].
\]

- Gauss \( P \) vs Laplace \( Q \)

![Graph showing witness function for Gauss and Laplace densities](image-url)
Function Showing Difference in Distributions

- **Maximum mean discrepancy**: smooth function for $P$ vs $Q$

  \[
  \text{MMD}(P, Q; F) := \sup_{f \in F} \mathbb{E}_P f(x) - \mathbb{E}_Q f(y). \]

- **Classical results**: $\text{MMD}(P, Q; F) = 0$ iff $P = Q$, when
  - $F =$ bounded continuous [Dudley, 2002]
  - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
  - $F =$ bounded Lipschitz (Earth mover’s distances) [Dudley, 2002]
Function Showing Difference in Distributions

- **Maximum mean discrepancy**: smooth function for $P$ vs $Q$

  $$
  \text{MMD}(P, Q; F) := \sup_{f \in F} \left[ E_P f(x) - E_Q f(y) \right].
  $$

- **Classical results**: $\text{MMD}(P, Q; F) = 0$ iff $P = Q$, when
  - $F =$ bounded continuous \[\text{[Dudley, 2002]}\]
  - $F =$ bounded variation 1 (Kolmogorov metric) \[\text{[Müller, 1997]}\]
  - $F =$ bounded Lipschitz (Earth mover’s distances) \[\text{[Dudley, 2002]}\]

- $\text{MMD}(P, Q; F) = 0$ iff $P = Q$ when $F =$ the unit ball in a characteristic \text{RKHS} $\mathcal{F}$ \[\text{Sriperumbudur et al. (2010), Gretton et al. (2012), Sejdinovic et al. (2013)}\]
The mean trick

The kernel trick: $\forall f \in \mathcal{F}$

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}} \quad \text{and} \quad \langle \varphi_{x_1}, \varphi_{x_2} \rangle_{\mathcal{F}} = k(x_1, x_2)$$

for positive definite $k(x, y)$.

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) = \langle f, \varphi_x \rangle_{\mathcal{F}} \quad f = \sum_{i=1}^{m} \alpha_i \varphi_{x_i}$$
The mean trick

The kernel trick: $\forall f \in \mathcal{F}$

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}} \quad \text{and} \quad \langle \varphi_{x_1}, \varphi_{x_2} \rangle_{\mathcal{F}} = k(x_1, x_2)$$

for positive definite $k(x, y)$.

The mean trick:

$$\mathbb{E}_P(f(X)) = \mathbb{E}_P[\langle \varphi_X, f \rangle_{\mathcal{F}}]$$

$$=: \langle \mu_P, f \rangle_{\mathcal{F}}$$

$\mu_P$ gives you expectations of all RKHS functions

When $k$ characteristic, then $\mu_P$ unique, e.g. Gauss, Laplace, ...
The mean trick

The kernel trick: \( \forall f \in \mathcal{F} \)

\[ f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}} \quad \text{and} \quad \langle \varphi_{x_1}, \varphi_{x_2} \rangle_{\mathcal{F}} = k(x_1, x_2) \]

for positive definite \( k(x, y) \).

The mean trick:

\[
\mathbb{E}_P(f(X)) = \mathbb{E}_P[\langle \varphi_X, f \rangle_{\mathcal{F}}] \\
= : \langle \mu_P, f \rangle_{\mathcal{F}}
\]

\( \mu_P \) gives you expectations of all RKHS functions

When \( k \) characteristic, then \( \mu_P \) unique, e.g. Gauss, Laplace, ...
• The (kernel) MMD:

\[
\text{MMD}^2(P, Q; F) = \left( \sup_{f \in F} \left[ \mathbb{E}_P f(x) - \mathbb{E}_Q f(y) \right] \right)^2
\]

 Witness \( f \) for Gauss and Laplace densities
Function view vs feature mean view

- The (kernel) MMD:

\[
\text{MMD}^2(P, Q; F) = \left( \sup_{f \in F} [E_P f(x) - E_Q f(y)] \right)^2
\]

use

\[
E_P(f(x)) = E_P [\langle \varphi, f \rangle_F] =: \langle \mu_P, f \rangle_F
\]
Function view vs feature mean view

- The (kernel) MMD:

\[
\text{MMD}^2(P, Q; F)
\]

\[
= \left( \sup_{f \in F} \left[ E_P f(x) - E_Q f(y) \right] \right)^2
\]

\[
= \left( \sup_{f \in F} \left[ \langle f, \mu_P - \mu_Q \rangle_F \right] \right)^2
\]

use

\[
E_P(f(x)) = E_P \left[ \langle \varphi_x, f \rangle_F \right]
\]

\[
= : \langle \mu_P, f \rangle_F
\]
Function view vs feature mean view

- The (kernel) MMD:

\[
\text{MMD}^2(P, Q; F) = \left( \sup_{f \in F} \left[ E_P f(x) - E_Q f(y) \right] \right)^2
\]

\[
= \left( \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F \right)^2
\]

\[
= \| \mu_P - \mu_Q \|^2_F
\]

use

\[
\| \theta \|_F = \sup_{f \in F} \langle f, \theta \rangle_F
\]

Function view and feature view equivalent
Function view vs feature mean view

- The (kernel) MMD:

\[
\text{MMD}^2(P, Q; F) = \left( \sup_{f \in F} [E_P f(x) - E_Q f(y)] \right)^2
\]

\[
= \left( \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F \right)^2
\]

\[
= \| \mu_P - \mu_Q \|_F^2
\]

\[
= E_P k(x, x') + E_Q k(y, y') - 2E_{P,Q} k(x, y)
\]

(a) = within distrib. similarity,

(b) = cross-distrib. similarity
Experiment: amplitude modulated signals

Samples from P

Samples from Q
Results: AM signals

\[ m = 10,000 \text{ and scaling } a = 0.5. \text{ Average over 4124 trials. Gaussian noise added.} \]

Linear-cost test and kernel choice approach from [Gretton et al., 2012b]
MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]
  Related to [Feuerverger, 1993] and [Székely and Rizzo, 2009, Székely et al., 2007]

\[
\left( \sup_f \left[ E_{P_{XY}} f - E_{P_X P_Y} f \right] \right)^2 = \sup_{\|f\| \leq 1} \langle f, \mu_{P_{XY}} - \mu_{P_X P_Y} \rangle^2_{\mathcal{F} \times \mathcal{G}}
\]

\[
= \| \mu_{P_{XY}} - \mu_{P_X P_Y} \|^2_{\mathcal{F} \times \mathcal{G}} := HSIC(P_{XY}, P_X P_Y)
\]
MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]
  Related to [Feuerverger, 1993] and [Székely and Rizzo, 2009, Székely et al., 2007]

\[
\left( \sup_f \left[ \mathbb{E}_{P_{XY}} f - \mathbb{E}_{P_X P_Y} f \right] \right)^2 = \sup_{\|f\| \leq 1} \langle f, \mu_{P_{XY}} - \mu_{P_X P_Y} \rangle_{\mathcal{F} \times \mathcal{G}}^2 \\
= \|\mu_{P_{XY}} - \mu_{P_X P_Y}\|_{\mathcal{F} \times \mathcal{G}}^2 =: HSIC(P_{XY}, P_X P_Y)
\]
MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]
  Related to [Feuerverger, 1993] and [Székely and Rizzo, 2009, Székely et al., 2007]

\[
\left( \sup_f \left[ \mathbb{E}_{P_{XY}} f - \mathbb{E}_{P_X P_Y} f \right] \right)^2 = \sup_{\|f\| \leq 1} \left\langle f, \mu_{P_{XY}} - \mu_{P_X P_Y} \right\rangle_{\mathcal{F} \times \mathcal{G}}^2 \\
= \|\mu_{P_{XY}} - \mu_{P_X P_Y}\|_{\mathcal{F} \times \mathcal{G}}^2 : = HSIC(P_{XY}, P_X P_Y)
\]

HSIC using expectations of kernels:

Define RKHS \( \mathcal{F} \) on \( \mathcal{X} \) with kernel \( k \), RKHS \( \mathcal{G} \) on \( \mathcal{Y} \) with kernel \( l \). Then

\[
HSIC(P_{XY}, P_X P_Y) \\
= E_{XY} E_{X'Y'} k(X, X')l(Y, Y') + E_X E_{X'} k(X, X')E_Y E_{Y'} l(Y, Y') \\
- 2E_{X'Y'} \left[ E_X k(X, X')E_Y l(Y, Y') \right].
\]
Experiment: dependence testing for translation

- **Translation example:** [NIPS07b] Canadian Hansard (agriculture)
- **5-line extracts,**
  - *k*-spectrum kernel, \( k = 10, \)
  - repetitions=300,
  - sample size 10
- **Empirical**
  - \( \text{MMD}(\mathbf{P}_{X Y}, \mathbf{P}_X \mathbf{P}_Y) : \)
    \[
    \frac{1}{n^2} (H K H \circ H L H)_{++}
    \]
- **\( k \)-spectrum kernel:** average Type II error 0 (\( \alpha = 0.05 \))
- **Bag of words kernel:** average Type II error 0.18
Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\text{COCO}(P; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} (\mathbb{E}_{x,y}[f(x)g(y)] - \mathbb{E}_x[f(x)]\mathbb{E}_y[g(y)])$$
Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\text{COCO} ( \mathbb{P}, \mathcal{F}, \mathcal{G} ) := \sup_{\|f\|_{\mathcal{F}} = 1, \|g\|_{\mathcal{G}} = 1} \left( \mathbb{E}_{x, y} [f(x)g(y)] - \mathbb{E}_x [f(x)] \mathbb{E}_y [g(y)] \right)$$
Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\text{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} (\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)])$$
Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

\[
\text{COCO}(\mathbb{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left( \mathbb{E}_{x, y}[f(x)g(y)] - \mathbb{E}_x[f(x)]\mathbb{E}_y[g(y)] \right)
\]
A more intuitive idea: **maximize covariance** of smooth mappings:

\[
\text{COCO}(P; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} (E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)])
\]
More functions revealing dependence

- Find the next best mappings
- (mappings orthogonal to first pair)
More functions revealing dependence

- Find the next best mappings
- (mappings orthogonal to first pair)
• Given $\gamma_i := \text{COCO}_i(z; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10] :

$$\text{HSIC}(z; \mathcal{F}, \mathcal{G}) := \sum_{i=1}^{n} \gamma_i^2$$
Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \text{COCO}_i(z; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10]:

$$\text{HSIC}(z; \mathcal{F}, \mathcal{G}) := \sum_{i=1}^{n} \gamma_i^2$$

- In limit of infinite samples:

$$\text{HSIC}(P; F, G) := \mathbf{E}_{x,x',y,y'}[k(x, x')l(y, y')] + \mathbf{E}_{x,x'}[k(x, x')]\mathbf{E}_{y,y'}[l(y, y')] - 2\mathbf{E}_{x,y}[\mathbf{E}_{x'}[k(x, x')]\mathbf{E}_{y'}[l(y, y')]]$$

- $x'$ an independent copy of $x$, $y'$ a copy of $y$

HSIC is identical to $\text{MMD}(P_{XY}, P_XP_Y)$
Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \text{COCO}_i(z; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10]:

$$\text{HSIC}(z; \mathcal{F}, \mathcal{G}) := \sum_{i=1}^{n} \gamma_i^2$$

- In limit of infinite samples:

$$\text{HSIC}(P; F, G) := \mathbf{E}_{x, x', y, y'} [k(x, x')l(y, y')] + \mathbf{E}_{x, x'} [k(x, x')]\mathbf{E}_{y, y'} [l(y, y')]$$

$$- 2\mathbf{E}_{x, y} [\mathbf{E}_{x'} [k(x, x')]\mathbf{E}_{y'} [l(y, y')]]$$

- $x'$ an independent copy of $x$, $y'$ a copy of $y$

HSIC is identical to $\text{MMD}(P_{XY}, P_X P_Y)$

Can we do better?
Kernel CCA

Before: smooth functions maximizing covariance (COCO)

Correlation: $-0.00$

Dependence witness, $X$

Dependence witness, $Y$

Correlation: $-0.90$  
COCO: $0.14$
Kernel CCA

Now: smooth functions maximizing correlation, Kernel CCA

Again, more dependence to be found:

- Ring-shaped density, third canonical correlate
**NOCCO: HS Norm of Normalized Cross Covariance**

- **NOCCO**: is Hiblert-Schmidt norm of correlation operator \([\text{Fukumizu et al., 2007}]\)

\[
\text{NOCCO} = \sum_i \text{KCC}_i^2
\]

\[
\widehat{\text{NOCCO}} := \text{tr}[R_y R_x], \quad R_x := \tilde{K}_x (\tilde{K}_x + n\epsilon_n I_n)^{-1}
\]
• **NOCCO:** is Hiblert-Schmidt norm of correlation operator \[\text{[Fukumizu et al., 2007]}\]

\[
\text{NOCCO} = \sum_i KCC_i^2
\]

\[
\widehat{\text{NOCCO}} := \text{tr}[R_y R_x], \quad R_x := \tilde{K}_x (\tilde{K}_x + n\epsilon_n I_n)^{-1}
\]

• For characteristic kernels: population NOCCO is mean-square contingency, indep. of RKHS

\[
\text{NOCCO} = \int \int_{\mathcal{X} \times \mathcal{Y}} \left( \frac{p_{xy}(x, y)}{p_x(x)p_y(y)} - 1 \right)^2 p_x(x)p_y(y) d\mu(x)d\mu(y).
\]

- \(\mu(x)\) and \(\mu(y)\) Lebesgue measures on \(\mathcal{X}\) and \(\mathcal{Y}\); \(P_{xy}\) absolutely continuous w.r.t. \(\mu(x) \times \mu(y)\), density \(p_{xy}\), marginal densities \(p_x\) and \(p_y\)
**NOCCO: HS Norm of Normalized Cross Covariance**

- **NOCCO**: is Hilbert-Schmidt norm of correlation operator \[ \text{NOCCO} = \sum_i KCC_i^2 \]

\[ \hat{\text{NOCCO}} := \text{tr} \left[ R_y R_x \right], \quad R_x := \tilde{K}_x (\tilde{K}_x + n\epsilon_n I_n)^{-1} \]

- For characteristic kernels: population NOCCO is mean-square contingency, indep. of RKHS

\[ \text{NOCCO} = \int \int_{\mathcal{X} \times \mathcal{Y}} \left( \frac{p_{xy}(x, y)}{p_x(x)p_y(y)} - 1 \right)^2 p_x(x)p_y(y)d\mu(x)d\mu(y). \]

- \( \mu(x) \) and \( \mu(y) \) Lebesgue measures on \( \mathcal{X} \) and \( \mathcal{Y} \); \( P_{xy} \) absolutely continuous w.r.t. \( \mu(x) \times \mu(y) \), density \( p_{xy} \), marginal densities \( p_x \) and \( p_y \)

- **Consistency result**: assume regularization \( \epsilon_n \) satisfies \( \epsilon_n \to 0 \) and \( \epsilon_n^3 n \to \infty \), Then \( \| \hat{V}_{xy} - V_{xy} \|_{HS} \overset{P}{\to} 0 \)
Energy Distance and the MMD
Energy distance and MMD

Distance between probability distributions:

**Energy distance:** \cite{Baringhaus_and_Franz_2004, Szekely_and_Rizzo_2004, Szekely_and_Rizzo_2005}

\[
D_E(P, Q) = \mathbb{E}_P \| X - X' \| + \mathbb{E}_Q \| Y - Y' \| - 2 \mathbb{E}_{P,Q} \| X - Y \|
\]

**Maximum mean discrepancy** \cite{Gretton_et_al_2007, Smola_et_al_2007, Gretton_et_al_2012a}

\[
\text{MMD}^2(P, Q; F) = \mathbb{E}_P k(X, X') + \mathbb{E}_Q k(Y, Y') - 2 \mathbb{E}_{P,Q} k(X, Y)
\]
Energy distance and MMD

Distance between probability distributions:

Energy distance: \cite{Baringhaus2004, Szekely2004, Szekely2005}

\[
D_E(P, Q) = \mathbb{E}_P \|X - X'\| + \mathbb{E}_Q \|Y - Y'\| - 2\mathbb{E}_{P, Q} \|X - Y\|
\]

Maximum mean discrepancy \cite{Gretton2007, Smola2007, Gretton2012a}

\[
\text{MMD}^2(P, Q; F) = \mathbb{E}_P k(X, X') + \mathbb{E}_Q k(Y, Y') - 2\mathbb{E}_{P, Q} k(X, Y)
\]

Energy distance is MMD with a particular kernel!
Distance covariance

\[ \mathcal{V}^2(X, Y) = E_{XY} E_{X'Y'} \left[ \|X, X'\| \|Y, Y'\| \right] + E_X E_{X'} \|X, X'\| E_Y E_{Y'} \|Y, Y'\| - 2E_{XY} \left[ E_{X'} \|X, X'\| E_{Y'} \|Y, Y'\| \right] \]

Hilbert-Schmidt Independence Criterion

Define RKHS \( \mathcal{F} \) on \( X \) with kernel \( k \), RKHS \( \mathcal{G} \) on \( Y \) with kernel \( l \). Then

\[ \text{HSIC}(P_{XY}, P_X P_Y) = E_{XY} E_{X'Y'} k(X, X') l(Y, Y') + E_X E_{X'} k(X, X') E_Y E_{Y'} l(Y, Y') - 2E_{X'Y'} \left[ E_X k(X, X') E_Y l(Y, Y') \right]. \]
Distance covariance and HSIC

Distance covariance \cite{Szekely and Rizzo, 2009, Szekely et al., 2007}

\[ \mathcal{V}^2(X, Y) = \mathbb{E}_{XY} \mathbb{E}_{X'Y'} \left[ \|X, X'\| \|Y, Y'\| \right] \]

\[ + \mathbb{E}_X \mathbb{E}_{X'} \|X, X'\| \mathbb{E}_Y \mathbb{E}_{Y'} \|Y, Y'\| \]

\[ - 2 \mathbb{E}_{XY} \left[ \mathbb{E}_{X'} \|X, X'\| \mathbb{E}_{Y'} \|Y, Y'\| \right] . \]

Hilbert-Schmidt Indepence Criterion \cite{Gretton et al., 2005, Smola et al., 2007, Gretton et al., 2008, Gretton and Gyorfi, 2010}

Define RKHS \( \mathcal{F} \) on \( \mathcal{X} \) with kernel \( k \), RKHS \( \mathcal{G} \) on \( \mathcal{Y} \) with kernel \( l \). Then

\[ \text{HSIC}(\mathbb{P}_{XY}, \mathbb{P}_X \mathbb{P}_Y) \]

\[ = \mathbb{E}_{XY} \mathbb{E}_{X'Y'} k(X, X')l(Y, Y') + \mathbb{E}_{X} \mathbb{E}_{X'} k(X, X') \mathbb{E}_{Y} \mathbb{E}_{Y'} l(Y, Y') \]

\[ - 2 \mathbb{E}_{X'Y'} \left[ \mathbb{E}_{X} k(X, X') \mathbb{E}_{Y} l(Y, Y') \right] . \]

Distance covariance is HSIC with particular kernels!
Semimetrics and Hilbert spaces

**Theorem** [Berg et al., 1984, Lemma 2.1, p. 74]

$\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semimetric on $\mathcal{X}$. Let $z_0 \in \mathcal{X}$, and denote

$$k_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

Then $k$ is **positive definite** (via Moore-Arnonsajn, defines a unique RKHS) iff $\rho$ is of negative type.

Call $k_\rho$ a **distance induced kernel**
Semimetrics and Hilbert spaces

Theorem [Berg et al., 1984, Lemma 2.1, p. 74]

\( \rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is a semimetric on \( \mathcal{X} \). Let \( z_0 \in \mathcal{X} \), and denote

\[
k_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').
\]

Then \( k \) is positive definite (via Moore-Arnonsajn, defines a unique RKHS) iff \( \rho \) is of negative type.

Call \( k_\rho \) a distance induced kernel

Special case: \( \mathcal{Z} \subseteq \mathbb{R}^d \) and \( \rho_q(z, z') = \|z - z'\|^q \). Then \( \rho_q \) is a valid semimetric of negative type for \( 0 < q \leq 2 \).
**Theorem** [Berg et al., 1984, Lemma 2.1, p. 74]

ρ : $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a semimetric on $\mathcal{X}$. Let $z_0 \in \mathcal{X}$, and denote

$$k_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

Then $k$ is positive definite (via Moore-Arnonsajn, defines a unique RKHS) iff $\rho$ is of negative type.

Call $k_\rho$ a distance induced kernel

**Special case:** $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\rho_q(z, z') = \|z - z'\|^q$. Then $\rho_q$ is a valid semimetric of negative type for $0 < q \leq 2$.

**Energy distance** is MMD with a distance induced kernel

**Distance covariance** is HSIC with distance induced kernels
Two-sample testing benchmark

Two-sample testing example in 1-D:

\[ P(X) \quad \text{VS} \quad Q(X) \]
Two-sample test, MMD with distance kernel

Obtain more powerful tests on this problem when $q \neq 1$ (exponent of distance)

Key:
- Gaussian kernel
- $q = 1$
- Best: $q = 1/3$
- Worst: $q = 2$
Lancaster (3-way) Interactions
Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?
Detecting a higher order interaction

• How to detect V-structures with pairwise weak (or nonexistent) dependence?
Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \independent Y$, $Y \independent Z$, $X \independent Z$

![Graph](image)

- $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$,
- $Z \mid X, Y \sim \text{sign}(XY) \exp\left(\frac{1}{\sqrt{2}}\right)$

Faithfulness violated here
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- **Consistent CI test:** $H_0 : X \perp\!\!\!\perp Y \mid Z$ [Fukumizu et al., 2008, Zhang et al., 2011], or
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- **Consistent CI test**: $H_0 : X \perp\!\!\!\perp Y \mid Z$ \[^{[Fukumizu et al., 2008, Zhang et al., 2011]}\], or

- **Factorisation test**: $H_0 : (X, Y) \perp\!\!\!\perp Z \lor (X, Z) \perp\!\!\!\perp Y \lor (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)
  - compute $p$-values for each of the marginal tests for $(Y, Z) \perp\!\!\!\perp X$, $(X, Z) \perp\!\!\!\perp Y$, or $(X, Y) \perp\!\!\!\perp Z$
  - apply Holm-Bonferroni (HB) sequentially rejective correction (Holm 1979)
V-structure Discovery (2)

- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- \( X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z \)

- \( X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1) \),
- \( Z_1 | X_1, Y_1 \sim \text{sign}(X_1Y_1)\text{Exp}\left(\frac{1}{\sqrt{2}}\right) \)
- \( X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1}) \) Faithfulness violated here
Figure 1: CI test for $X \perp Y | Z$ from Zhang et al (2011), and a factorisation test with a HB correction, $n = 500$
[Bahadur (1961); Lancaster (1969)] **Interaction measure** of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that **vanishes** whenever \(P\) can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- **\(D = 2\)**: \(\Delta_L P = P_{XY} - P_X P_Y\)
[Bahadur (1961); Lancaster (1969)] **Interaction measure** of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that **vanishes** whenever \(P\) can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- **\(D = 2\)**: \(\Delta_L P = P_{XY} - P_X P_Y\)
- **\(D = 3\)**: \(\Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z\)
Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] Interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- \(D = 2\) : \(\Delta_L P = P_{XY} - P_X P_Y\)
- \(D = 3\) : \(\Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z\)
[Bahadur (1961); Lancaster (1969)] Interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- **\(D = 2\)**: \(\Delta LP = P_{XY} - P_X P_Y\)
- **\(D = 3\)**: \(\Delta LP = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z\)

\[\Delta LP = 0\]

\(P_{XYZ}\)

\begin{align*}
\text{Case of } P_X & \perp \perp P_{YZ} \\
\text{\(X\)} & \text{\(Y\)} & \text{\(Z\)}
\end{align*}
[Bahadur (1961); Lancaster (1969)] Interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that \textbf{vanishes} whenever \(P\) can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- \(D = 2: \quad \Delta_L P = P_{XY} - P_X P_Y\)
- \(D = 3: \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z\)

\[(X, Y) \perp\!
\iff\!\!\perp Z \lor (X, Z) \perp\!
\iff\!\!\perp Y \lor (Y, Z) \perp\!
\iff\!\!\perp X \Rightarrow \Delta_L P = 0.\]

...so what might be missed?
Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure $\Delta P$ that vanishes whenever $P$ can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2$ : $\Delta_L P = P_{XY} - P_X P_Y$
- $D = 3$ : $\Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$

$\Delta_L P = 0 \nRightarrow (X, Y) \perp\!
\perp Z \lor (X, Z) \perp\!
\perp Y \lor (Y, Z) \perp\!
\perp X$

**Example:**

<table>
<thead>
<tr>
<th>$P(0, 0, 0)$</th>
<th>$P(0, 0, 1)$</th>
<th>$P(1, 0, 0)$</th>
<th>$P(1, 0, 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$P(0, 1, 0)$</td>
<td>$P(0, 1, 1)$</td>
<td>$P(1, 1, 0)$</td>
<td>$P(1, 1, 1)$</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>
A Test using Lancaster Measure

- Test statistic is empirical estimate of \( \| \mu_\kappa (\Delta_L P) \|_{H_\kappa}^2 \), where 
  \( \kappa = k \otimes l \otimes m \):

\[
\| \mu_\kappa (P_{XYZ} - P_{XY} P_{Z} - \cdots) \|_{H_\kappa}^2 = \\
\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{H_\kappa} - 2 \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY} P_{Z} \rangle_{H_\kappa} \cdots
\]
Inner Product Estimators

<table>
<thead>
<tr>
<th>$\nu \backslash \nu'$</th>
<th>$P_{XYZ}$</th>
<th>$P_{XY} P_Z$</th>
<th>$P_{XZ} P_Y$</th>
<th>$P_{YZ} P_X$</th>
<th>$P_X P_Y P_Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{XYZ}$</td>
<td>$(K \circ L \circ M)_{++}$</td>
<td>$((K \circ L) M)_{++}$</td>
<td>$((K \circ M) L)_{++}$</td>
<td>$((M \circ L) K)_{++}$</td>
<td>$tr(K_+ \circ L_+ \circ M_+)$</td>
</tr>
<tr>
<td>$P_{XY} P_Z$</td>
<td>$(K \circ L)<em>{++} M</em>{++}$</td>
<td>$(MKL)_{++}$</td>
<td>$(KLM)_{++}$</td>
<td>$(KL)<em>{++} M</em>{++}$</td>
<td></td>
</tr>
<tr>
<td>$P_{XZ} P_Y$</td>
<td>$(K \circ M)<em>{++} L</em>{++}$</td>
<td>$(KML)_{++}$</td>
<td>$(KM)<em>{++} L</em>{++}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{YZ} P_X$</td>
<td></td>
<td>$(L \circ M)<em>{++} K</em>{++}$</td>
<td>$(LM)<em>{++} K</em>{++}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_X P_Y P_Z$</td>
<td></td>
<td></td>
<td></td>
<td>$K_{++} L_{++} M_{++}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: $V$-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$
Inner Product Estimators

<table>
<thead>
<tr>
<th>$\nu \backslash \nu'$</th>
<th>$P_{XYZ}$</th>
<th>$P_{XY}P_{Z}$</th>
<th>$P_{XZ}P_{Y}$</th>
<th>$P_{YZ}P_{X}$</th>
<th>$P_{X}P_{Y}P_{Z}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{XYZ}$</td>
<td>$(K \circ L \circ M)_{++}$</td>
<td>$((K \circ L) M)_{++}$</td>
<td>$((K \circ M) L)_{++}$</td>
<td>$((M \circ L) K)_{++}$</td>
<td>$tr(K_{+} \circ L_{+} \circ M_{+})$</td>
</tr>
<tr>
<td>$P_{XY}P_{Z}$</td>
<td>$(K \circ L)<em>{++} M</em>{++}$</td>
<td>$(MKL)_{++}$</td>
<td>$(KLM)_{++}$</td>
<td>$(KL)<em>{++} M</em>{++}$</td>
<td></td>
</tr>
<tr>
<td>$P_{XZ}P_{Y}$</td>
<td>$(K \circ M)<em>{++} L</em>{++}$</td>
<td>$(KML)_{++}$</td>
<td>$(KM)<em>{++} L</em>{++}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{YZ}P_{X}$</td>
<td></td>
<td>$(L \circ M)<em>{++} K</em>{++}$</td>
<td>$(LM)<em>{++} K</em>{++}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{X}P_{Y}P_{Z}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$K_{++} L_{++} M_{++}$</td>
</tr>
</tbody>
</table>

Table 2: $V$-statistic estimators of $\langle \mu_{K\nu}, \mu_{K\nu}' \rangle_{H_{K}}$

$$\|\mu_{K}(\Delta_{LP})\|^{2}_{H_{K}} = \frac{1}{n^{2}} \left( HKH \circ HLH \circ HMH \right)_{++}.$$ 

Empirical joint central moment in the feature space
Example A: factorisation tests

Figure 2: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with HB correction); Test for $X \perp Y | Z$ from Zhang et al (2011), $n = 500$
Example B: Joint dependence can be easier to detect

- \( X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \)
  \[
  \begin{align*}
  Z_1 = \begin{cases}
  X_1^2 + \epsilon, & \text{w.p. } 1/3, \\
  Y_1^2 + \epsilon, & \text{w.p. } 1/3, \\
  X_1Y_1 + \epsilon, & \text{w.p. } 1/3,
  \end{cases}
  \end{align*}
  \]
  where \( \epsilon \sim \mathcal{N}(0,0.1^2) \).

- \( X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{p-1}) \)

- dependence of \( Z \) on pair \((X, Y)\) is stronger than on \( X \) and \( Y \) individually

- Satisfies faithfulness
Example B: factorisation tests

Figure 3: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with HB correction); Test for $X \perp Y | Z$ from Zhang et al (2011), $n = 500$
Interaction for \( D \geq 4 \)

- Interaction measure valid for all \( D \) (Streitberg, 1990):

\[
\Delta_S P = \sum_{\pi} (-1)^{\left|\pi\right|-1} (\left|\pi\right| - 1)! J_{\pi} P
\]

- For a partition \( \pi \), \( J_{\pi} \) associates to the joint the corresponding factorisation, e.g.,

\[
J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}.
\]
Interaction for $D \geq 4$

- Interaction measure valid for all $D$ (Streitberg, 1990):

$$\Delta_SP = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_{\pi}P$$

- For a partition $\pi$, $J_{\pi}$ associates to the joint the corresponding factorisation, e.g.,

$$J_{13|2|4}P = P_{X_1X_3}P_{X_2}P_{X_4}.$$
Interaction for $D \geq 4$

- Interaction measure valid for all $D$ (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi| - 1} (|\pi| - 1)! J_{\pi} P$$

- For a partition $\pi$, $J_{\pi}$ associates to the joint the corresponding factorisation, e.g.,

$$J_{13|2|4}P = PX_1X_3PX_2PX_4.$$

**Joint central moments** (Lancaster interaction)

vs.

**Joint cumulants** (Streitberg interaction)
Total independence test

- Total independence test:
  \( H_0 : P_{XYZ} = P_X P_Y P_Z \) vs. \( H_1 : P_{XYZ} \neq P_X P_Y P_Z \)
Total independence test

- **Total independence test:**
  \[ H_0 : P_{XYZ} = P_X P_Y P_Z \text{ vs. } H_1 : P_{XYZ} \neq P_X P_Y P_Z \]

- For \((X_1, \ldots, X_D) \sim P_X\), and \(\kappa = \bigotimes_{i=1}^D k^{(i)}\):
  \[
  \mu_\kappa \left( \hat{P}_X - \prod_{i=1}^D \hat{P}_{X_i} \right) \bigg|_{\Delta_{tot} \hat{P}} \bigg| = \frac{1}{n^2} \sum_{a=1}^n \sum_{b=1}^n \prod_{i=1}^D K_{ab}^{(i)} - \frac{2}{n^{D+1}} \sum_{a=1}^n \prod_{i=1}^D \sum_{b=1}^n K_{ab}^{(i)} + \frac{1}{n^{2D}} \prod_{i=1}^D \sum_{a=1}^n \sum_{b=1}^n K_{ab}^{(i)}.
  \]

- Coincides with the test proposed by Kankainen (1995) using empirical characteristic functions: similar relationship to that between dCov and HSIC (DS et al, 2013)
Figure 4: Total independence: $\Delta_{tot} \hat{P}$ vs. $\Delta_L \hat{P}$, $n = 500$
Outline

- **An RKHS metric** on the space of probability measures
  - Distance between means in space of features (RKHS)
  - Nonparametric two-sample test...
  - ...for (almost!) any data type: strings, images, graphs, groups (rotation matrices), semigroups (histograms),...

- **Dependence detection**
  - Covariance and Correlation in feature space

- **Relation with energy distance and distance covariance**

- **Interactions with three (or more) variables**
Co-authors

- From UCL:
  - Luca Baldasssarre
  - Steffen Grunewalder
  - Guy Lever
  - Sam Patterson
  - Massimiliano Pontil
  - Dino Sejdinovic

- External:
  - Karsten Borgwardt, MPI
  - Wicher Bergsma, LSE
  - Kenji Fukumizu, ISM
  - Zaid Harchaoui, INRIA
  - Bernhard Schoelkopf, MPI
  - Alex Smola, CMU/Google
  - Le Song, Georgia Tech
  - Bharath Sriperumbudur, Cambridge
Selected references

Characteristic kernels and mean embeddings:


Two-sample, independence, conditional independence tests:


Energy distance, relation to kernel distances


Three way interaction

Selected references (continued)

**Conditional mean embedding, RKHS-valued regression:**

**Kernel Bayes rule:**
Local departures from the null

What is a hard testing problem?
Local departures from the null

What is a hard testing problem?

- **First version:** for fixed $m$, “closer” $P$ and $Q$ have higher Type II error
Local departures from the null

What is a hard testing problem?

- As $m$ increases, distinguish “closer” $P$ and $Q$ with fixed Type II error
Local departures from the null

What is a hard testing problem?

- As \( m \) increases, distinguish “closer” \( P \) and \( Q \) with fixed Type II error

- Example: \( f_P \) and \( f_Q \) probability densities, \( f_Q = f_P + \delta g \), where \( \delta \in \mathbb{R} \), \( g \) some fixed function such that \( f_Q \) is a valid density
  - If \( \delta \sim m^{-1/2} \), Type II error approaches a constant
More general local departures from null

- **Example:** $f_P$ and $f_Q$ probability densities, $f_Q = f_P + \delta g$, where $\delta \in \mathbb{R}$, $g$ some fixed function such that $f_Q$ is a valid density.

![Graphs showing $f_P$ and $f_Q$](attachment:image.png)
Local departures from the null

What is a hard testing problem?

- As we see more samples \( m \), distinguish “closer” \( P \) and \( Q \) with same Type II error

- Example: \( f_P \) and \( f_Q \) probability densities, \( f_Q = f_P + \delta g \), where \( \delta \in \mathbb{R} \), \( g \) some fixed function such that \( f_Q \) is a valid density
  - If \( \delta \sim m^{-1/2} \), Type II error approaches a constant

- ...but other choices also possible – how to characterize them all?
Local departures from the null

What is a hard testing problem?

- As we see more samples \( m \), distinguish “closer” \( P \) and \( Q \) with same Type II error

- **Example:** \( f_P \) and \( f_Q \) probability densities, \( f_Q = f_P + \delta g \), where \( \delta \in \mathbb{R} \), \( g \) some *fixed* function such that \( f_Q \) is a valid density
  - If \( \delta \sim m^{-1/2} \), Type II error approaches a constant

- ...but **other choices also possible** – how to characterize them all?

General characterization of local departures from \( \mathcal{H}_0 \):

- Write \( \mu_Q = \mu_P + g_m \), where \( g_m \in \mathcal{F} \) chosen such that \( \mu_P + g_m \) a valid distribution embedding

- Minimum distinguishable distance \([\text{JMLR12}]\)

\[
\|g_m\|_{\mathcal{F}} = c m^{-1/2}
\]
More general local departures from null

- More advanced example of a local departure from the null
- Recall: $\mu_Q = \mu_P + g_m$, and $\|g_m\|_F = c m^{-1/2}$
Kernels vs kernels

- How does MMD relate to Parzen density estimate? [Anderson et al., 1994]

\[ \hat{f}_P(x) = \frac{1}{m} \sum_{i=1}^{m} \kappa(x_i - x), \] where \( \kappa \) satisfies \( \int_{\mathcal{X}} \kappa(x) \, dx = 1 \) and \( \kappa(x) \geq 0 \).
Kernels vs kernels

- How does MMD relate to Parzen density estimate? [Anderson et al., 1994]

\[ \hat{f}_P(x) = \frac{1}{m} \sum_{i=1}^{m} \kappa(x_i - x), \text{ where } \kappa \text{ satisfies } \int \kappa(x) \, dx = 1 \text{ and } \kappa(x) \geq 0. \]

- \textbf{L}_2 \text{ distance} between Parzen density estimates:

\[
D_2(\hat{f}_P, \hat{f}_Q)^2 = \int \left[ \frac{1}{m} \sum_{i=1}^{m} \kappa(x_i - z) - \frac{1}{m} \sum_{i=1}^{m} \kappa(y_i - z) \right]^2 \, dz
\]

\[
= \frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i - x_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(y_i - y_j) - \frac{2}{m^2} \sum_{i,j=1}^{m} k(x_i - y_j),
\]

where \( k(x - y) = \int \kappa(x - z) \kappa(y - z) \, dz \)
Kernels vs kernels

- How does MMD relate to Parzen density estimate? [Anderson et al., 1994]

\[ \hat{f}_P(x) = \frac{1}{m} \sum_{i=1}^{m} \kappa(x_i - x), \quad \text{where } \kappa \text{ satisfies } \int \kappa(x) \, dx = 1 \text{ and } \kappa(x) \geq 0. \]

- \( L_2 \) distance between Parzen density estimates:

\[
D_2(\hat{f}_P, \hat{f}_Q)^2 = \int \left[ \frac{1}{m} \sum_{i=1}^{m} \kappa(x_i - z) - \frac{1}{m} \sum_{i=1}^{m} \kappa(y_i - z) \right]^2 dz
= \frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i - x_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(y_i - y_j) - \frac{2}{m^2} \sum_{i,j=1}^{m} k(x_i - y_j),
\]

where \( k(x - y) = \int \kappa(x - z) \kappa(y - z) \, dz \)

- \( f_Q = f_P + \delta g \), minimum distance to discriminate \( f_P \) from \( f_Q \) is

\[ \delta = (m)^{-1/2} h_m^{-d/2}, \quad \text{where } h_m \text{ is width of } \kappa. \]
Kernel two-sample tests for big data, optimal kernel choice
Quadratic time estimate of MMD

\[ \text{MMD}^2 = \| \mu_P - \mu_Q \|^2_F = E_P k(x, x') + E_Q k(y, y') - 2E_{P,Q} k(x, y) \]
Quadratic time estimate of MMD

\[ \text{MMD}^2 = \|\mu_P - \mu_Q\|_F^2 = E_P k(x, x') + E_Q k(y, y') - 2E_{P,Q} k(x, y) \]

Given i.i.d. \( X := \{x_1, \ldots, x_m\} \) and \( Y := \{y_1, \ldots, y_m\} \) from \( P, Q \), respectively:

The earlier estimate: (quadratic time)

\[ \hat{E}_P k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j\neq i} k(x_i, x_j) \]
Quadratic time estimate of MMD

\[ \text{MMD}^2 = \| \mu_P - \mu_Q \|^2_F = \mathbb{E}_P k(x, x') + \mathbb{E}_Q k(y, y') - 2\mathbb{E}_{P,Q} k(x, y) \]

Given i.i.d. \( X := \{x_1, \ldots, x_m\} \) and \( Y := \{y_1, \ldots, y_m\} \) from \( P, Q \), respectively:

The earlier estimate: (quadratic time)

\[ \hat{\mathbb{E}}_P k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} k(x_i, x_j) \]

New, linear time estimate:

\[ \hat{\mathbb{E}}_P k(x, x') = \frac{2}{m} \left[ k(x_1, x_2) + k(x_3, x_4) + \ldots \right] \]

\[ = \frac{2}{m} \sum_{i=1}^{m/2} k(x_{2i-1}, x_{2i}) \]
Linear time MMD

Shorter expression with explicit $k$ dependence:

$$\text{MMD}^2 =: \eta_k(p, q) = \mathbb{E}_{xx', yy'} h_k(x, x', y, y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') - k(x', y),$$

and $v := [x, x', y, y']$. 
Shorter expression with explicit $k$ dependence:

$$\text{MMD}^2 =: \eta_k(p, q) = \mathbb{E}_{xx', yy'} h_k(x, x', y, y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') - k(x', y),$$

and $v := [x, x', y, y']$.

The linear time estimate again:

$$\hat{\eta}_k = \frac{2}{m} \sum_{i=1}^{m/2} h_k(v_i),$$

where $v_i := [x_{2i-1}, x_{2i}, y_{2i-1}, y_{2i}]$ and

$$h_k(v_i) := k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(x_{2i}, y_{2i-1})$$
Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given $m$, hence...
- ...a much less powerful test for a given $m$
Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

• Much higher variance for a given $m$, hence...
• ...a much less powerful test for a given $m$

Advantages of the linear time MMD vs quadratic time MMD

• Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of $\chi^2$)
• Both test statistic and threshold computable in $O(m)$, with storage $O(1)$.
• Given unlimited data, a given Type II error can be attained with less computation
Asymptotics of linear time MMD

By central limit theorem,

\[ m^{1/2} (\tilde{\eta}_k - \eta_k(p, q)) \overset{D}{\to} \mathcal{N}(0, 2\sigma_k^2) \]

- assuming \( 0 < \mathbb{E}(h_k^2) < \infty \) (true for bounded \( k \))
- \( \sigma_k^2 = \mathbb{E}_v h_k^2(v) - [\mathbb{E}_v(h_k(v))]^2 \).
Hypothesis test

Hypothesis test of asymptotic level $\alpha$:

$$t_{k, \alpha} = m^{-1/2} \sigma_k \sqrt{2\Phi^{-1}(1 - \alpha)}$$

where $\Phi^{-1}$ is inverse CDF of $\mathcal{N}(0, 1)$. 

Null distribution, linear time $\widehat{\text{MMD}}^2 = \hat{\eta}_k$
Type II error

Null vs alternative distribution, $P(\hat{\eta}_k)$

Type II error

$\eta_k(p, q)$
The best kernel: minimizes Type II error

Type II error: $\tilde{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p, q) > 0$.

Prob. of a Type II error:

$$P(\tilde{\eta}_k < t_{k,\alpha}) = \Phi \left( \Phi^{-1}(1 - \alpha) - \frac{\eta_k(p, q) \sqrt{m}}{\sigma_k \sqrt{2}} \right)$$

where $\Phi$ is a Normal CDF.
The best kernel: minimizes Type II error

Type II error: $\tilde{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p, q) > 0$.

Prob. of a Type II error:

$$P(\tilde{\eta}_k < t_{k,\alpha}) = \Phi \left( \Phi^{-1}(1 - \alpha) - \frac{\eta_k(p, q)\sqrt{m}}{\sigma_k\sqrt{2}} \right)$$

where $\Phi$ is a Normal CDF.

Since $\Phi$ monotonic, best kernel choice to minimize Type II error prob. is:

$$k_* = \arg \max_{k \in \mathcal{K}} \frac{\eta_k(p, q)}{\sigma_k},$$

where $\mathcal{K}$ is the family of kernels under consideration.
Learning the best kernel in a family

Define the family of kernels as follows:

\[ \mathcal{K} := \left\{ k : k = \sum_{u=1}^{d} \beta_u k_u, \|\beta\|_1 = D, \beta_u \geq 0, \forall u \in \{1, \ldots, d\} \right\}. \]

Properties: if at least one \( \beta_u > 0 \)

- all \( k \in \mathcal{K} \) are valid kernels,
- If all \( k_u \) characteristic then \( k \) characteristic
The squared MMD becomes

$$
\eta_k(p, q) = \| \mu_k(p) - \mu_k(q) \|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p, q),
$$

where $\eta_u(p, q) := \mathbb{E}_v h_u(v)$.
Test statistic

The squared MMD becomes

\[ \eta_k(p, q) = \|\mu_k(p) - \mu_k(q)\|_F^2 = \sum_{u=1}^{d} \beta_u \eta_u(p, q), \]

where \( \eta_u(p, q) := \mathbb{E}_v h_u(v) \).

Denote:

- \( \beta = (\beta_1, \beta_2, \ldots, \beta_d)^\top \in \mathbb{R}^d \),
- \( h = (h_1, h_2, \ldots, h_d)^\top \in \mathbb{R}^d \),
- \( h_u(x, x', y, y') = k_u(x, x') + k_u(y, y') - k_u(x, y') - k_u(x', y) \)
- \( \eta = \mathbb{E}_v(h) = (\eta_1, \eta_2, \ldots, \eta_d)^\top \in \mathbb{R}^d \).

Quantities for test:

\[ \eta_k(p, q) = \mathbb{E}(\beta^\top h) = \beta^\top \eta \quad \sigma_k^2 := \beta^\top \text{cov}(h)\beta. \]
Optimization of ratio $\eta_k(p, q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta}$$

$$\hat{\sigma}_{k, \lambda} = \sqrt{\beta^\top \left( \hat{Q} + \lambda_m I \right) \beta},$$

$\hat{Q}$ is empirical estimate of cov$(h)$.

**Note:** $\hat{\eta}_k, \hat{\sigma}_{k, \lambda}$ computed on training data, vs $\tilde{\eta}_k, \tilde{\sigma}_k$ on data to be tested (why?)
Optimization of ratio $\eta_k(p, q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \quad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top (\hat{Q} + \lambda m I) \beta},$$

$\hat{Q}$ is empirical estimate of cov($h$).

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\tilde{\eta}_k, \tilde{\sigma}_k$ on data to be tested (why?)

Objective:

$$\hat{\beta}^* = \arg\max_{\beta \succeq 0} \hat{\eta}_k(p, q)\hat{\sigma}_{k,\lambda}^{-1}$$

$$= \arg\max_{\beta \succeq 0} \left( \beta^\top \hat{\eta} \right) \left( \beta^\top (\hat{Q} + \lambda m I) \beta \right)^{-1/2}$$

$$=: \alpha(\beta; \hat{\eta}, \hat{Q})$$
Optimization of ratio $\eta_k(p, q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry
Then there exists $\beta \succeq 0$ s.t. $\alpha(\beta; \hat{\eta}, \hat{Q}) > 0$.
Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$
Optimization of ratio $\eta_k(p, q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\beta \succeq 0$ s.t. $\alpha(\beta; \hat{\eta}, \hat{Q}) > 0$.

Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\hat{\beta}^* = \arg \max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q})$.

Quadratic program:

$$\min \{ \beta^\top (\hat{Q} + \lambda_m I) \beta : \beta^\top \hat{\eta} = 1, \beta \succeq 0 \}$$
Optimization of ratio $\eta_k(p, q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\beta \succeq 0$ s.t. $\alpha(\beta; \hat{\eta}, \hat{Q}) > 0$.

Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\hat{\beta}^* = \arg\max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q})$.

Quadratic program:

$$\min\{\beta^\top (\hat{Q} + \lambda_m I) \beta : \beta^\top \hat{\eta} = 1, \beta \succeq 0\}$$

What if $\hat{\eta}$ has no positive entries?
1. Split the data into testing and training.

2. On the training data:
   (a) Compute $\hat{\eta}_u$ for all $k_u \in \mathcal{K}$
   (b) If at least one $\hat{\eta}_u > 0$, solve the QP to get $\beta^*$, else choose random kernel from $\mathcal{K}$

3. On the test data:
   (a) Compute $\tilde{\eta}_{k^*}$ using $k^* = \sum_{u=1}^{d} \beta^* k_u$
   (b) Compute test threshold $\tilde{t}_{\alpha,k^*}$ using $\tilde{\sigma}_{k^*}$

4. Reject null if $\tilde{\eta}_{k^*} > \tilde{t}_{\alpha,k^*}$
Convergence bounds

Assume bounded kernel, $\sigma_k$, bounded away from 0. If $\lambda_m = \Theta(m^{-1/3})$ then

$$
\left| \sup_{k \in K} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in K} \eta_k \sigma_k^{-1} \right| = O_P \left( m^{-1/3} \right).
$$
Convergence bounds

Assume bounded kernel, $\sigma_k$, bounded away from 0. If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in K} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in K} \eta_k \sigma_k^{-1} \right| = O_P \left( m^{-1/3} \right).$$

Idea:

$$\left| \sup_{k \in K} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in K} \eta_k \sigma_k^{-1} \right| \leq \sup_{k \in K} \left| \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \eta_k \sigma_k^{-1} \right| + \sup_{k \in K} \left| \eta_k \sigma_k^{-1} - \eta_k \sigma_k^{-1} \right|$$

$$\leq \frac{\sqrt{d}}{D \sqrt{\lambda_m}} \left( C_1 \sup_{k \in K} |\hat{\eta}_k - \eta_k| + C_2 \sup_{k \in K} |\hat{\sigma}_{k,\lambda} - \sigma_{k,\lambda}| \right) + C_3 D^2 \lambda_m,$$
Experiments
Competing approaches

- Median heuristic
- Max. MMD: choose $k_u \in \mathcal{K}$ with the largest $\hat{\eta}_u$
  - same as maximizing $\beta^T \hat{\eta}$ subject to $\|\beta\|_1 \leq 1$
- $\ell_2$ statistic: maximize $\beta^T \hat{\eta}$ subject to $\|\beta\|_2 \leq 1$
- Cross validation on training set

Also compare with:

- Single kernel that maximizes ratio $\eta_k(p, q)\sigma_k^{-1}$
Blobs: data

**Difficult problems**: lengthscale of the *difference* in distributions not the same as that of the distributions.
Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

We distinguish a field of Gaussian blobs with different covariances.

Ratio $\varepsilon = 3.2$ of largest to smallest eigenvalues of blobs in $q$. 

![Blob data](image-url)
Parameters: $m = 10,000$ (for training and test). Ratio $\varepsilon$ of largest to smallest eigenvalues of blobs in $q$. Results are average over 617 trials.
Blobs: results

Optimize ratio $\eta_k(p, q)\sigma_k^{-1}$
Maximize $\eta_k(p, q)$ with $\beta$ constraint
Feature selection: data

Idea: no single best kernel.
Each of the $k_u$ are univariate (along a single coordinate)
Idea: no single best kernel.
Each of the $k_u$ are univariate (along a single coordinate)
Feature selection: results

$m = 10,000$, average over 5000 trials
Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

$$u(t) = \sin(\omega_c t) [a \ s(t) + l]$$

- $\omega_c$: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset
Amplitude modulated signals

Given an audio signal \( s(t) \), an amplitude modulated signal can be defined

\[
u(t) = \sin(\omega_c t) [a s(t) + l]
\]

- \( \omega_c \): carrier frequency
- \( a = 0.2 \) is signal scaling, \( l = 2 \) is offset

Two amplitude modulated signals from same artist (in this case, Magnetic Fields).

- Music sampled at 8KHz \( \text{(very low)} \)
- Carrier frequency is 24kHz
- AM signal observed at 120kHz
- Samples are extracts of length \( N = 1000 \), approx. 0.01 sec \( \text{(very short)} \).
- Total dataset size is 30,000 samples from each of \( p, q \).
Amplitude modulated signals

Samples from P

Samples from Q
Results: AM signals

\[ m = 10,000 \text{ (for training and test)} \] and scaling \( a = 0.5 \). Average over 4124 trials. Gaussian noise added.
Observations on kernel choice

- It is possible to choose the best kernel for a kernel two-sample test

- Kernel choice matters for “difficult” problems, where the distributions differ on a lengthscale different to that of the data.

- Ongoing work:
  - quadratic time statistic
  - avoid training/test split
References


