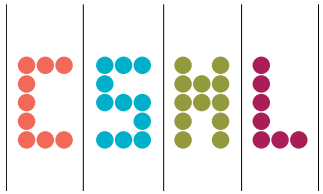
Hypothesis Testing with Kernel Embeddings

kernel selection for large-scale tests and equivalence to energy distance

Dino Sejdinovic (Gatsby Unit, CSML, UCL)

joint work with:
Arthur Gretton (Gatsby Unit, CSML, UCL), Bharath Sriperumbudur (StatsLab, Cambridge),
Heiko Strathmann (CSML, UCL), Sivaraman Balakrishnan (LTI, CMU),
Massimiliano Pontil (CSML, UCL), Kenji Fukumizu (ISM, Tokyo)

12 November 2012



Overview

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

Two-sample and independence tests

- **Two-sample test**: Given $\{Z^{(i)}\}_{i=1}^{n_z} \overset{i.i.d.}{\sim} P$, and $\{W^{(i)}\}_{i=1}^{n_w} \overset{i.i.d.}{\sim} Q$,
  - H_0: $P = Q$
  - H_A: $P \neq Q$

Two-sample and independence tests

- **Two-sample test**: Given $\{Z^{(i)}\}_{i=1}^{n_z} \overset{i.i.d.}{\sim} P$, and $\{W^{(i)}\}_{i=1}^{n_w} \overset{i.i.d.}{\sim} Q$,
  - H_0: $P = Q$
  - H_A: $P \neq Q$
- **Independence test**: Given $\left\{\left(X^{(i)}, Y^{(i)}\right)\right\}_{i=1}^{m} \overset{i.i.d.}{\sim} P_{XY}$,
  - H_0: $P_{XY} = P_X P_Y$
  - H_A: $P_{XY} \neq P_X P_Y$

Two-sample and independence tests

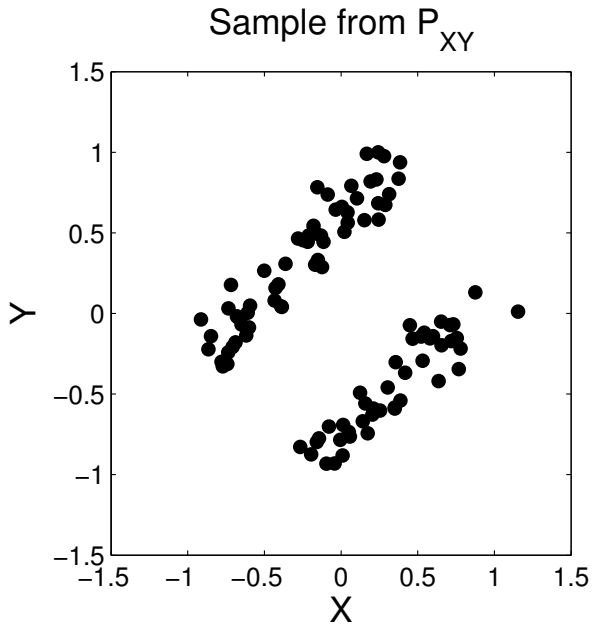
- **Two-sample test**: Given $\{Z^{(i)}\}_{i=1}^{n_z} \overset{i.i.d.}{\sim} P$, and $\{W^{(i)}\}_{i=1}^{n_w} \overset{i.i.d.}{\sim} Q$,
  - H_0: $P = Q$
  - H_A: $P \neq Q$
- **Independence test**: Given $\left\{\left(X^{(i)}, Y^{(i)}\right)\right\}_{i=1}^{m} \overset{i.i.d.}{\sim} P_{XY}$,
  - H_0: $P_{XY} = P_X P_Y$
  - H_A: $P_{XY} \neq P_X P_Y$
- high-dimensions

Two-sample and independence tests

- **Two-sample test**: Given $\{Z^{(i)}\}_{i=1}^{n_z} \overset{i.i.d.}{\sim} P$, and $\{W^{(i)}\}_{i=1}^{n_w} \overset{i.i.d.}{\sim} Q$,
  - H_0: $P = Q$
  - H_A: $P \neq Q$
- **Independence test**: Given $\left\{\left(X^{(i)}, Y^{(i)}\right)\right\}_{i=1}^{m} \overset{i.i.d.}{\sim} P_{XY}$,
  - H_0: $P_{XY} = P_X P_Y$
  - H_A: $P_{XY} \neq P_X P_Y$
- high-dimensions
- non-Euclidean / structured domains

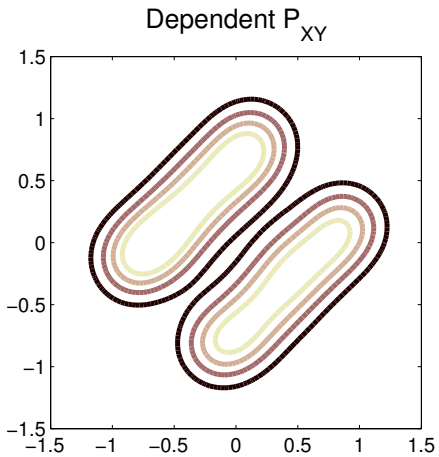
Motivating question

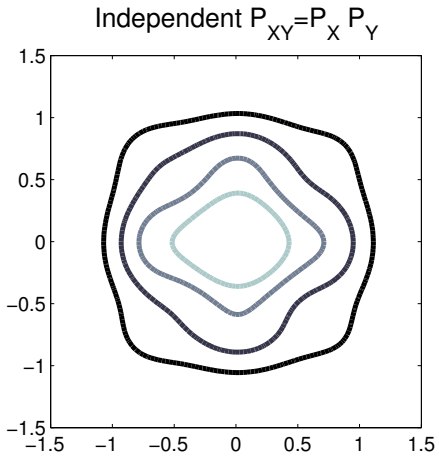
- How do you detect dependence...
- ... in a Euclidean space?



dependent, but uncorrelated

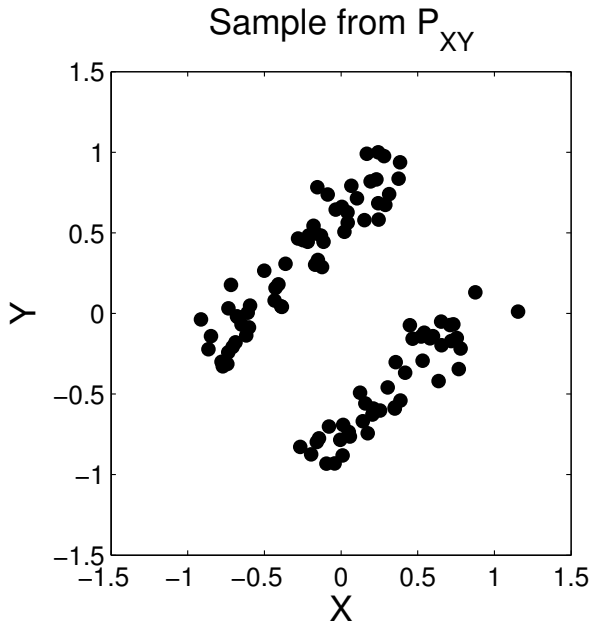
?



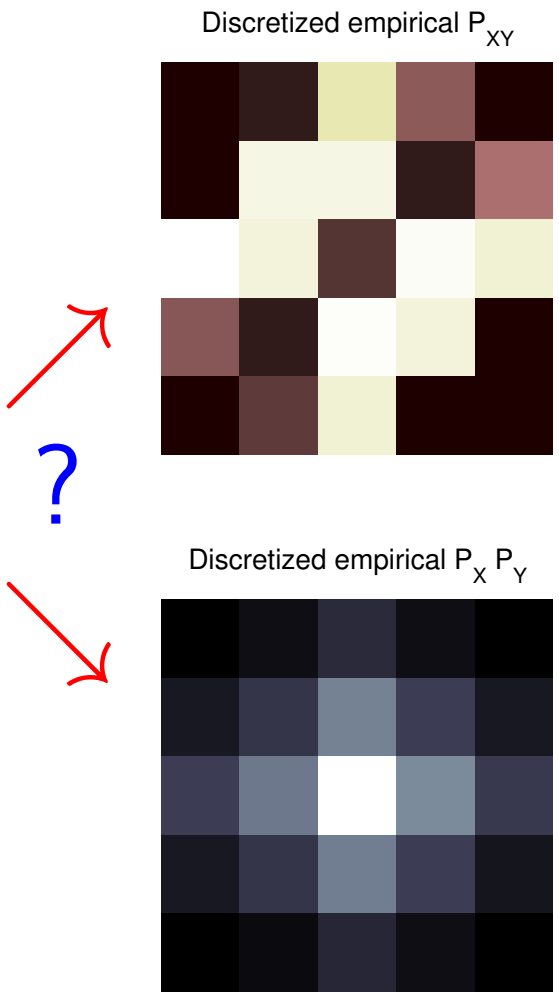


Motivating question

- How do you detect dependence...
- ... in a Euclidean space?

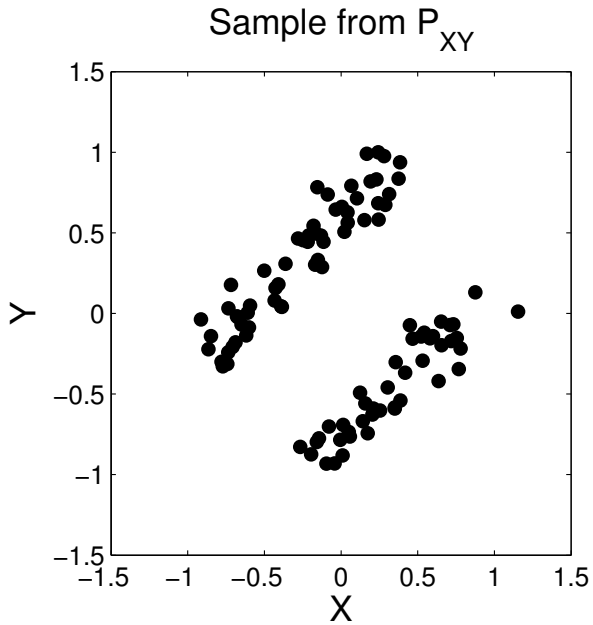


dependent, but uncorrelated



Motivating question

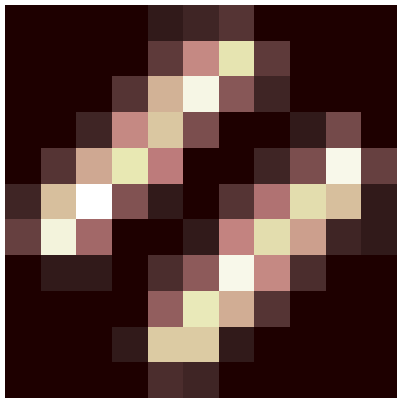
- How do you detect dependence...
- ... in a Euclidean space?



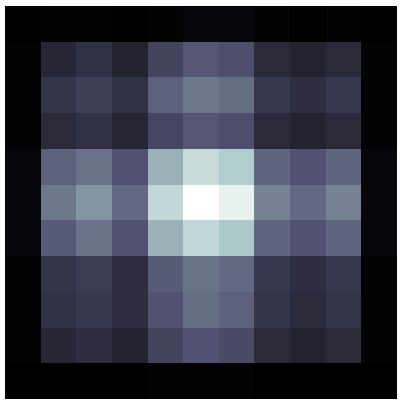
dependent, but uncorrelated

?

Discretized empirical P_{XY}



Discretized empirical P_X P_Y



Motivating question

- How do you detect dependence...
- ... in a Euclidean space?
- Problem: fails even in "low" dimensions: too few points per bin!

Motivating question

- How do you detect dependence...
- ... in a Euclidean space?
- Problem: fails even in "low" dimensions: too few points per bin!

Task: representing and comparing distributions in high dimensions

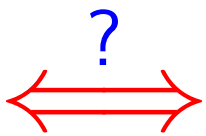
Motivating question

- How do you detect dependence...
- ... in a non-Euclidean / structured domain?

Motivating question

- How do you detect dependence...
- ... in a non-Euclidean / structured domain?

...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...

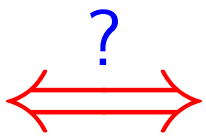


...il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement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

Motivating question

- How do you detect dependence...
- ... in a non-Euclidean / structured domain?

...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...



...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

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

RKHS

Definition (**RKHS**)

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall z \in \mathcal{Z}, \quad k(\cdot, z) \in \mathcal{H}$, and
2. $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space* (RKHS).

RKHS

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall z \in \mathcal{Z}, \quad k(\cdot, z) \in \mathcal{H}$, and
2. $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space* (RKHS).

- Evaluation functionals are continuous.
- Norm convergence implies pointwise convergence.

RKHS

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall z \in \mathcal{Z}, \quad k(\cdot, z) \in \mathcal{H}$, and
2. $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space* (RKHS).

- Evaluation functionals are continuous.
- Norm convergence implies pointwise convergence.
- L^2 is not an RKHS as $\delta_z \notin L^2$

Feature map

- Extract features: $z \mapsto (\phi_1(z), \dots \phi_s(z)) \in \mathbb{R}^s$, and work with kernel $k(z, z') = \sum_{i=1}^{s} \phi_i(z)\phi_i(z')$ (inner product in the feature space)

Feature map

- Extract features: $z \mapsto (\phi_1(z), \dots \phi_s(z)) \in \mathbb{R}^s$, and work with kernel $k(z, z') = \sum_{i=1}^{s} \phi_i(z)\phi_i(z')$ (inner product in the feature space)

Theorem (Moore-Aronszajn)

For every symmetric, positive semi-definite function (kernel) $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, there is a unique associated RKHS $\mathcal{H}_k$ of real-valued functions on $\mathcal{Z}$ with reproducing kernel k.

- The map $\varphi : \mathcal{Z} \to \mathcal{H}_k$, $\varphi : z \mapsto k(\cdot, z)$ is called the canonical feature map or the Aronszajn map of k.

Kernel Embeedding

Definition (Kernel embedding)

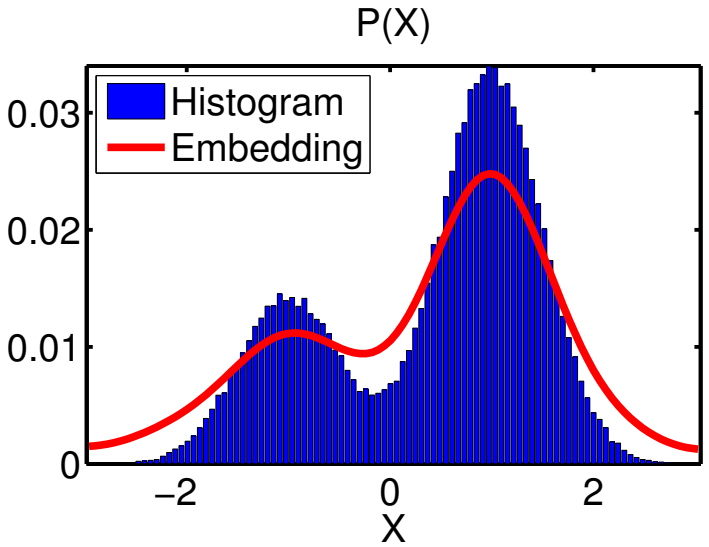
Let k be a kernel on $\mathcal{Z}$, and $P \in \mathcal{M}_+^1(\mathcal{Z})$ a probability measure. The *kernel embedding* of P into the RKHS $\mathcal{H}_k$ is $\mu_k(P) \in \mathcal{H}_k$ such that $\int f(z) dP(z) = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k}$ **for all** $f \in \mathcal{H}_k$.

Kernel Embeedding

Definition (Kernel embedding)

Let k be a kernel on $\mathcal{Z}$, and $P \in \mathcal{M}_+^1(\mathcal{Z})$ a probability measure. The *kernel embedding* of P into the RKHS $\mathcal{H}_k$ is $\mu_k(P) \in \mathcal{H}_k$ such that $\int f(z) dP(z) = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k}$ **for all** $f \in \mathcal{H}_k$.

- Alternatively, can be defined by the Bochner integral $\mu_k(P) = \int k(\cdot, z)\, dP(z)$ ["Expected canonical feature"]



Existence of Kernel Embedding

Proposition

$\mu_k(P)$ exists for all $P \in \mathcal{M}_+^1(\mathcal{Z})$ iff k is a bounded function on $\mathcal{Z} \times \mathcal{Z}$.

Existence of Kernel Embedding

Proposition

$\mu_k(P)$ exists for all $P \in \mathcal{M}_+^1(\mathcal{Z})$ iff k is a bounded function on $\mathcal{Z} \times \mathcal{Z}$.

Denote:

$$\mathcal{M}_k^\theta(\mathcal{Z}) = \left\{ \nu \in \mathcal{M}(\mathcal{Z}) \, : \, \int k^\theta(z,z) \, d|\nu|(z) < \infty \right\}.$$

- Consequence of the Riesz representation theorem: kernel embedding $\mu_k(\nu)$ is well defined $\forall \nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$.

Maximum Mean Discrepancy

- $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a kernel on $\mathcal{Z}$, with RKHS $\mathcal{H}_k$; P, Q two probability measure on $\mathcal{Z}$:
- **Maximum Mean Discrepancy (MMD)** between P and Q:

$$\begin{aligned} \gamma_k(P, Q) &= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} \\ &= [\mathbb{E}_{ZZ'} k(Z, Z') + \mathbb{E}_{WW'} k(W, W') - 2\mathbb{E}_{ZW} k(Z, W)]^{1/2} \end{aligned}$$

Maximum Mean Discrepancy

- $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a kernel on $\mathcal{Z}$, with RKHS $\mathcal{H}_k$; P, Q two probability measure on $\mathcal{Z}$:
- **Maximum Mean Discrepancy (MMD)** between P and Q:

$$\begin{aligned}\gamma_k(P, Q) &= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} \\ &= \left[\mathbb{E}_{ZZ'} k(Z, Z') + \mathbb{E}_{WW'} k(W, W') - 2\mathbb{E}_{ZW} k(Z, W)\right]^{1/2}\end{aligned}$$

- A polynomial kernel $k(z, z') = \left(1 + z^\top z'\right)^p$ captures the difference in first p moments only

Maximum Mean Discrepancy

- $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a kernel on $\mathcal{Z}$, with RKHS $\mathcal{H}_k$; P, Q two probability measure on $\mathcal{Z}$:
- **Maximum Mean Discrepancy (MMD)** between P and Q:

$$\begin{aligned}\gamma_k(P, Q) &= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} \\ &= [\mathbb{E}_{ZZ'} k(Z, Z') + \mathbb{E}_{WW'} k(W, W') - 2\mathbb{E}_{ZW} k(Z, W)]^{1/2}\end{aligned}$$

- A polynomial kernel $k(z, z') = \left(1 + z^\top z'\right)^p$ captures the difference in first p moments only
- For a certain family of kernels (**characteristic**): $\gamma_k(P, Q) = 0$ if and only if $P = Q$.

Maximum Mean Discrepancy

- $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a kernel on $\mathcal{Z}$, with RKHS $\mathcal{H}_k$; P, Q two probability measure on $\mathcal{Z}$:
- **Maximum Mean Discrepancy (MMD)** between P and Q:

$$\begin{aligned}\gamma_k(P, Q) &= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} \\ &= [\mathbb{E}_{ZZ'} k(Z, Z') + \mathbb{E}_{WW'} k(W, W') - 2\mathbb{E}_{ZW} k(Z, W)]^{1/2}\end{aligned}$$

- A polynomial kernel $k(z, z') = \left(1 + z^\top z'\right)^p$ captures the difference in first p moments only
- For a certain family of kernels (**characteristic**): $\gamma_k(P, Q) = 0$ if and only if $P = Q$.
- Gaussian, Laplacian, inverse multiquadratics...

MMD for independence: HSIC

- $k_{\mathcal{X}}$ a kernel on $\mathcal{X}$, $k_{\mathcal{Y}}$ a kernel on $\mathcal{Y}$; then $k = k_{\mathcal{X}} k_{\mathcal{Y}}$ is a valid kernel on $\mathcal{X} \times \mathcal{Y}$ with RKHS $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$.
- **Hilbert-Schmidt Independence Criterion** between X and Y:

$$\begin{aligned} HSIC^2(X, Y; k_{\mathcal{X}}, k_{\mathcal{Y}}) &= \|\mu_k(P_{XY}) - \mu_k(P_X P_Y)\|^2_{\mathcal{H}_k} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'} k_{\mathcal{X}}(X, X') k_{\mathcal{Y}}(Y, Y') \\ &\quad + \mathbb{E}_X \mathbb{E}_{X'} k_{\mathcal{X}}(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} k_{\mathcal{Y}}(Y, Y') \\ &\quad - 2\mathbb{E}_{X'Y'} \left[\mathbb{E}_X k_{\mathcal{X}}(X, X') \mathbb{E}_Y k_{\mathcal{Y}}(Y, Y') \right]. \end{aligned}$$

- Gretton et al (2005, 2008); Smola et al (2007); Zhang et al (2011); Gretton et al (2012)

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

V-statistic

- Write $\eta_k := \gamma_k^2 = \mathbb{E}_{ZZ'} k(Z, Z') + \mathbb{E}_{WW'} k(W, W') - 2\mathbb{E}_{ZW} k(Z, W)$.
- Given i.i.d. samples $\mathbf{z} = \{z_i\}_{i=1}^m \sim P$ and $\mathbf{w} = \{w_i\}_{i=1}^m \sim Q$, the empirical V-statistic estimate of η_k is given by:

$$\begin{aligned}
\hat{\eta}_{k,V}(\mathbf{z}, \mathbf{w}) &= \gamma_k^2 \left(\frac{1}{m} \sum_{i=1}^m \delta_{z_i}, \frac{1}{m} \sum_{j=1}^m \delta_{w_j} \right) \\
&= \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(z_i, z_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(w_i, w_j) \\
&\quad - \frac{2}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(z_i, w_j).
\end{aligned}$$

A quadratic time estimate.

Asymptotics of quadratic time MMD

Theorem (Gretton et al, 2009)

Let k be a kernel on $\mathcal{Z}$, and let $\mathbf{z} = \{z_i\}_{i=1}^m$ and $\mathbf{w} = \{w_i\}_{i=1}^m$ be two i.i.d. samples from $P \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^1(\mathcal{Z})$. Then

$$\frac{m}{2}\hat{\eta}_{k,V}(\mathbf{z},\mathbf{w}) \quad \rightsquigarrow \quad \sum_{i=1}^{\infty} \lambda_i N_i^2,$$

where $N_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $i \in \mathbb{N}$, and $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of the operator $S_{\tilde{k}_P} : L_P^2(\mathcal{Z}) \to L_P^2(\mathcal{Z})$, given by:

$$S_{\tilde{k}_P} g(z) \quad = \quad \int_{\mathcal{Z}} \tilde{k}_P(z,w) g(w)\, dP(w).$$

Linear time estimate of MMD

Alternative expression:

$$\gamma_k^2(P, Q) = \mathbb{E}_{XX'YY'} h_k(X, X', Y, Y') =: \mathbb{E}_V h_k(V),$$

where

$$h_k(X, X', Y, Y') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y),$$

and $V := [X,\ X',\ Y,\ Y'] \sim P \times P \times Q \times Q$.

Linear time estimate of MMD

Alternative expression:

$$\gamma_k^2(P, Q) = \mathbb{E}_{XX'YY'} h_k(X, X', Y, Y') =: \mathbb{E}_V h_k(V),$$

where

$$h_k(X, X', Y, Y') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y),$$

and $V := [X, X', Y, Y'] \sim P \times P \times Q \times Q$.

A linear time estimate: Given i.i.d. samples $\{v_i\}_{i=1}^{m/2}$, with $v_i = [x_{2i-1}, x_{2i}, y_{2i-1}, y_{2i}]$

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

Linear time estimate of MMD

Alternative expression:

$$\gamma_k^2(P, Q) = \mathbb{E}_{XX'YY'} h_k(X, X', Y, Y') =: \mathbb{E}_V h_k(V),$$

where

$$h_k(X, X', Y, Y') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y),$$

and $V := [X, X', Y, Y'] \sim P \times P \times Q \times Q$.

A linear time estimate: Given i.i.d. samples $\{v_i\}_{i=1}^{m/2}$, with $v_i = [x_{2i-1}, x_{2i}, y_{2i-1}, y_{2i}]$

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

An empirical average of i.i.d. (quadruples of) samples.

Asymptotics of linear time MMD

By central limit theorem,

$$\sqrt{\frac{m}{2}}\left(\hat{\eta}_{k,L} - \eta_k\right) \rightsquigarrow \mathcal{N}(0, \sigma_k^2)$$

- assuming $0 < \mathbb{E}(h_k^2) < \infty$ (always true for bounded k)
- $\sigma_k^2 = \mathbb{E}_V h_k^2(V) - [\mathbb{E}_V(h_k(V))]^2 = \mathrm{var}(h_k(V))$

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

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

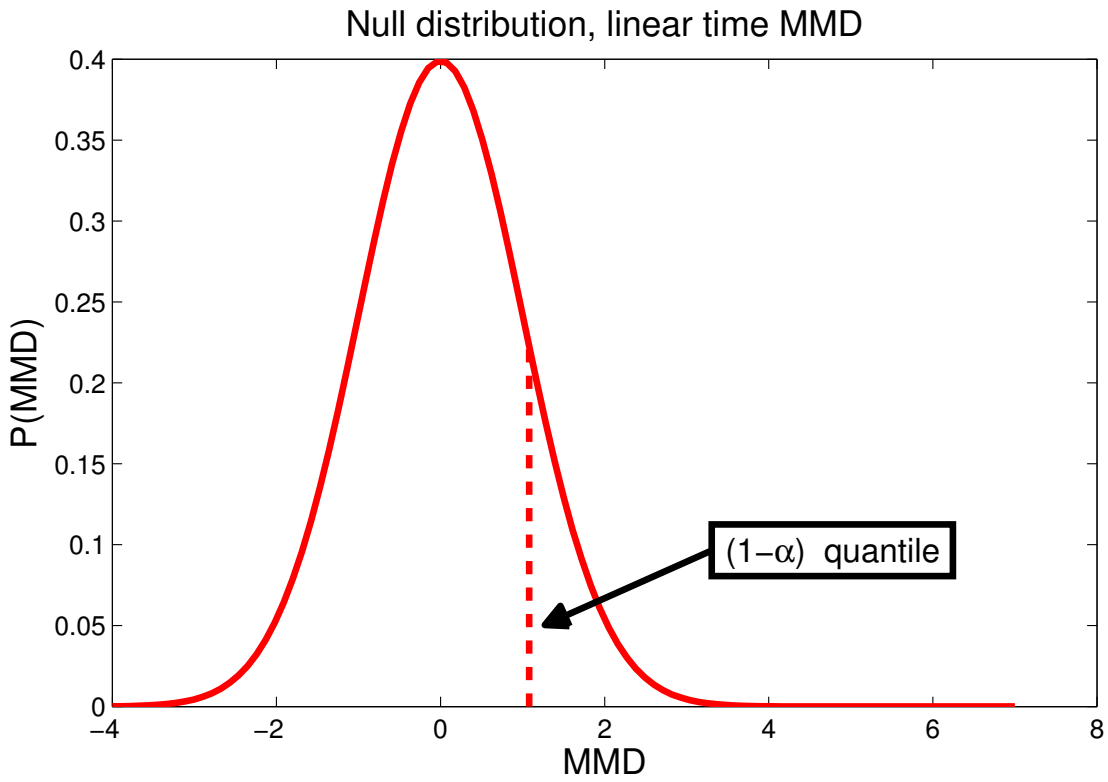
Advantages of the linear time MMD vs quadratic time MMD

- Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of χ^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

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

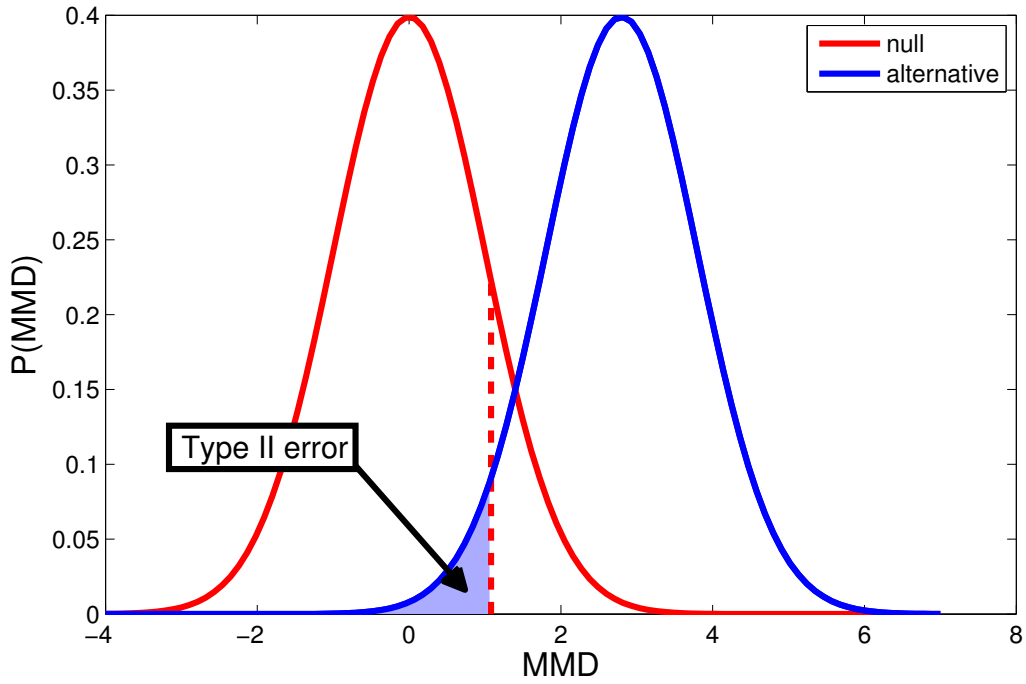
Testing threshold



Under null, $\eta_k = 0$, and thus: $\hat{\eta}_{k,L} \approx \mathcal{N}(0, \frac{2}{m}\sigma_k^2)$ leads to the threshold for an asymptotic level α:

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

Type II error



- Type II error: $\eta_k(P, Q) > 0$ and $\hat{\eta}_{k,L}$ falls below the threshold $t_{k,\alpha}$:

$$\begin{aligned}\mathbb{P}(\hat{\eta}_{k,L} < t_{k,\alpha}) &\approx \mathbb{P}(\eta_k + \sqrt{\frac{2}{m}}\sigma_k Z < \sqrt{\frac{2}{m}}\sigma_k \Phi^{-1}(1-\alpha)) \\ &= \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{\frac{m}{2}}\frac{\eta_k}{\sigma_k}\right)\end{aligned}$$

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

The best kernel: minimizes Type II error

Since Φ monotonic, the kernel which minimizes the asymptotic Type II error prob. is:

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

where $\mathcal{K}$ is the family of kernels under consideration.

The best kernel: minimizes Type II error

Since Φ monotonic, the kernel which minimizes the asymptotic Type II error prob. is:

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

where $\mathcal{K}$ is the family of kernels under consideration.

- We only have estimates of η_k and σ_k.
  - Will the optimization using these esimates be consistent?
  - Over what families of kernels can we perform such optimization?

Learning the best kernel in a family

Define the family of kernels as follows: for *base* kernels $\{k_u\}_{u=1}^d$

$$\mathcal{K} := \left\{ k \ : \ k = \sum_{u=1}^{d} \beta_u k_u, \ \sum_{u=1}^{d} \beta_u = 1, \ \beta_u \geq 0, \ \forall u \right\}.$$

Properties:

- all $k \in \mathcal{K}$ are valid kernels,
- if all k_u characteristic then every $k \in \mathcal{K}$ is characteristic

Test statistic

The squared MMD η_k becomes

$$\eta_k(P, Q) = \sum_{u=1}^{d} \beta_u \eta_u(P, Q),$$

where we denoted $\eta_u := \eta_{k_u}$.

Denote:

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

Then

$$\eta_k(P, Q) = \mathbb{E}(\beta^\top \mathbf{h}) = \beta^\top \eta.$$

Linear time estimates of variance

To implement both **kernel selection** and the **hypothesis test**, we need an empirical variance estimate.

$$\sigma_u^2 = \mathrm{var}(h_u).$$

Linear time estimates of variance

To implement both **kernel selection** and the **hypothesis test**, we need an empirical variance estimate.

$$\sigma_u^2 = \mathrm{var}(h_u).$$

Case of $k = \sum_{u=1}^{d} \beta_u k_u \in \mathcal{K}$:

$$\sigma_k^2 := \beta^\top Q \beta, \qquad Q = \mathrm{cov}(\mathbf{h}).$$

Linear time estimates of variance

To implement both **kernel selection** and the **hypothesis test**, we need an empirical variance estimate.

$$\sigma_u^2 = \mathrm{var}(h_u).$$

Case of $k = \sum_{u=1}^{d} \beta_u k_u \in \mathcal{K}$:

$$\sigma_k^2 := \beta^\top Q \beta, \qquad Q = \mathrm{cov}(\mathbf{h}).$$

Linear time estimates:

$$\hat{\sigma}_u^2 = \frac{4}{m} \sum_{i=1}^{m/4} (h_u(v_{2i-1}) - h_u(v_{2i}))^2.$$

Linear time estimates of variance

To implement both **kernel selection** and the **hypothesis test**, we need an empirical variance estimate.

$$\sigma_u^2 = \mathrm{var}(h_u).$$

Case of $k = \sum_{u=1}^{d} \beta_u k_u \in \mathcal{K}$:

$$\sigma_k^2 := \beta^\top Q \beta, \qquad Q = \mathrm{cov}(\mathbf{h}).$$

Linear time estimates:

$$\hat{\sigma}_u^2 = \frac{4}{m} \sum_{i=1}^{m/4} (h_u(v_{2i-1}) - h_u(v_{2i}))^2.$$

$$\hat{Q}_{uu'} = \frac{4}{m} \sum_{i=1}^{m/4} \left[h_u(v_{2i-1}) - h_u(v_{2i})\right] \left[h_{u'}(v_{2i-1}) - h_{u'}(v_{2i})\right].$$

Surrogate criterion

Define

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

Surrogate criterion

Define

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

Objective:

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

Surrogate criterion

Define

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

Objective:

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

Note: $\hat{\eta}_k, \hat{\sigma}_k$ used in optimization are computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ computed on data to be tested (**makes kernel choice independent of test data**)

Consistency

Theorem

If k_u is bounded $\forall u \in \{1, \ldots, d\}$ and $\lambda_m = \Theta(m^{-1/3})$, then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P\left(m^{-1/3}\right), \quad \text{and} \quad \hat{k}_* \xrightarrow{P} k_*.$$

Optimization procedure

- Assume: $\hat{\eta}$ has at least one positive entry
  - Then there exists $\beta \succeq 0$ s.t. $\beta^\top \hat{\eta} > 0$ (criterion is non-negative at optimality).

Optimization procedure

- Assume: $\hat{\eta}$ has at least one positive entry
  - Then there exists $\beta \succeq 0$ s.t. $\beta^\top \hat{\eta} > 0$ (criterion is non-negative at optimality).

Then, we can solve an easier problem (a quadratic program):

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

Optimization procedure

- Assume: $\hat{\eta}$ has at least one positive entry
  - Then there exists $\beta \succeq 0$ s.t. $\beta^\top \hat{\eta} > 0$ (criterion is non-negative at optimality).

Then, we can solve an easier problem (a quadratic program):

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

What if $\hat{\eta}$ has no positive entries? (all training data empirical MMDs on base kernels are ≤ 0)

Optimization procedure

- Assume: $\hat{\eta}$ has at least one positive entry
  - Then there exists $\beta \succeq 0$ s.t. $\beta^\top \hat{\eta} > 0$ (criterion is non-negative at optimality).

Then, we can solve an easier problem (a quadratic program):

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

What if $\hat{\eta}$ has no positive entries? (all training data empirical MMDs on base kernels are ≤ 0)
Cost: linear in the number of samples, quadratic in the number of kernels.

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

Competing approaches

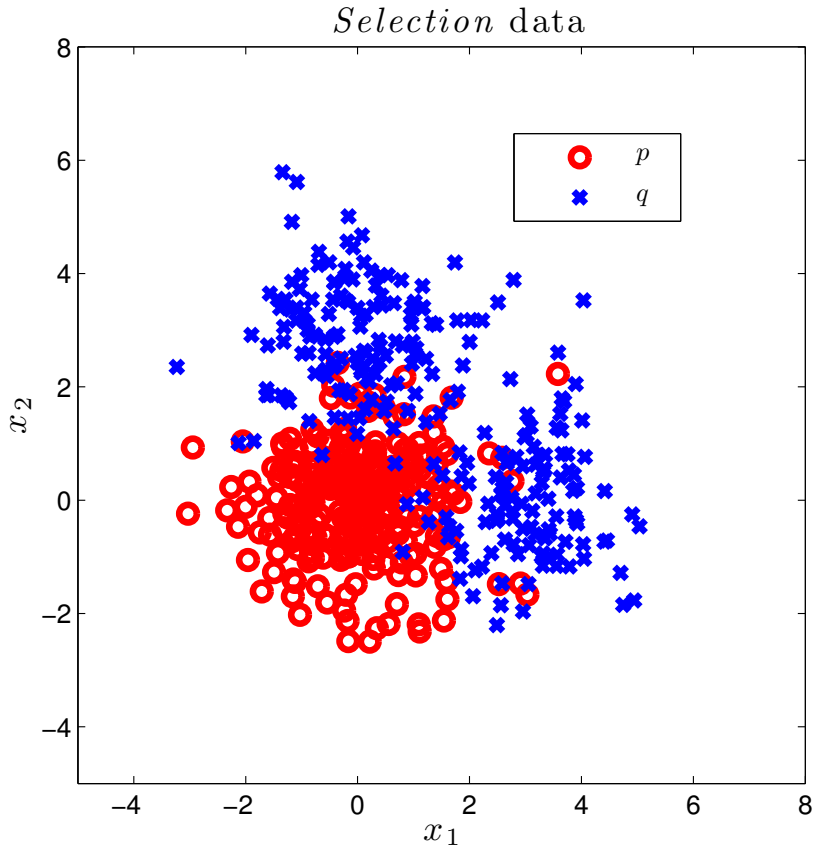
- Median heuristic
- max-mmd: choose $k_u \in \mathcal{K}$ with the largest $\hat{\eta}_u$
  - same as maximizing $\beta^\top \hat{\eta}$ subject to $\|\beta\|_1 = 1$
- ℓ_2 statistic: maximize $\beta^\top \hat{\eta}$ subject to $\|\beta\|_2 = 1$
- Cross validation on training set

Also compare with:

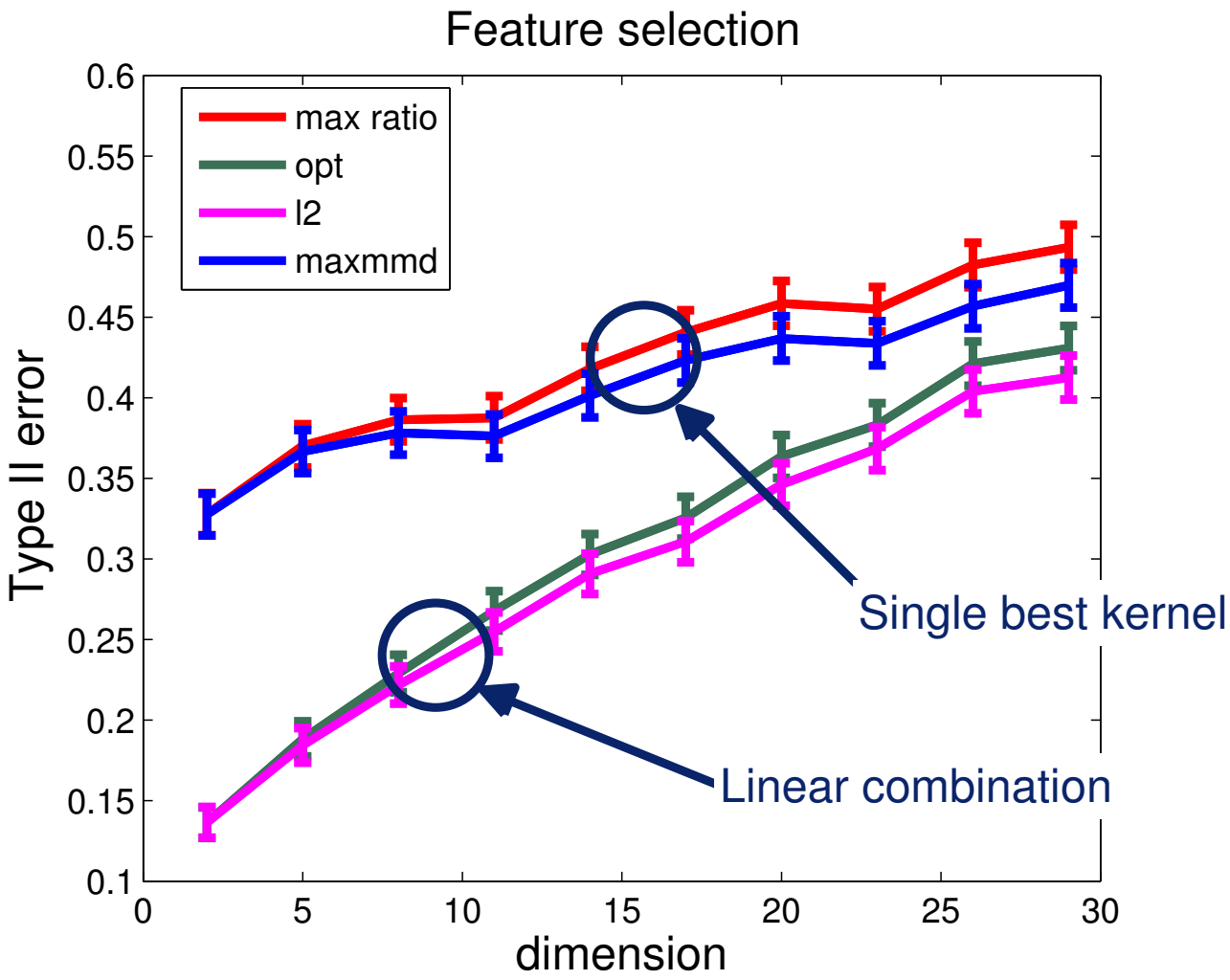
- max-ratio: single kernel k_u that maximizes $\hat{\eta}_u \hat{\sigma}_{u,\lambda}^{-1}$

Experiment 1: Feature selection

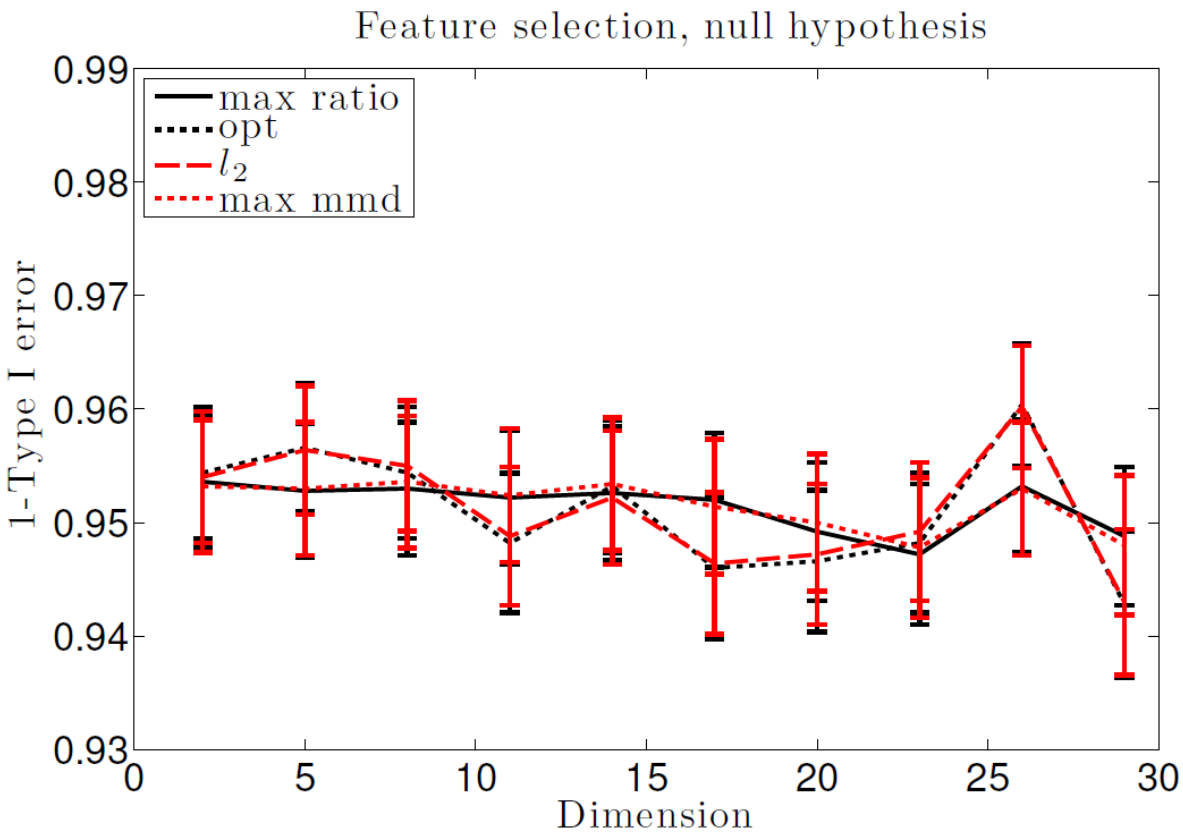
Idea: in this experiment, no single best kernel.
Each of the k_u are univariate (along a single coordinate)



Feature selection: Type II error



Feature selection: Type I error



Experiment 2: Grid-mixtures

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

Experiment 2: Grid-mixtures

Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.
We distinguish grids of Gaussian blobs with different covariances.

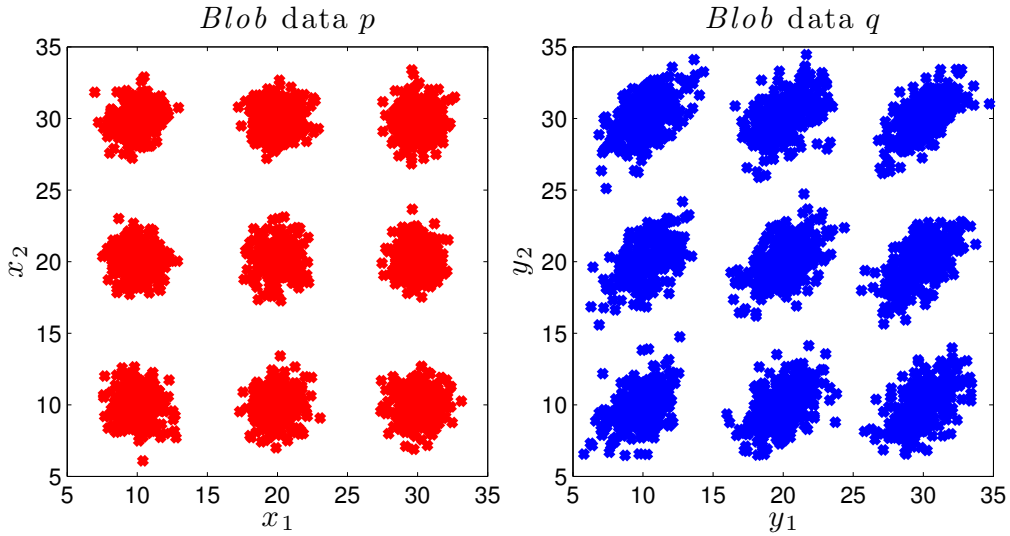


Figure: $\ell = 3$ of blobs per dimension, ratio $\varepsilon = 3.2$ of largest-to-smallest eigenvalues of blobs in Q.

Grid-mixtures: Type II error

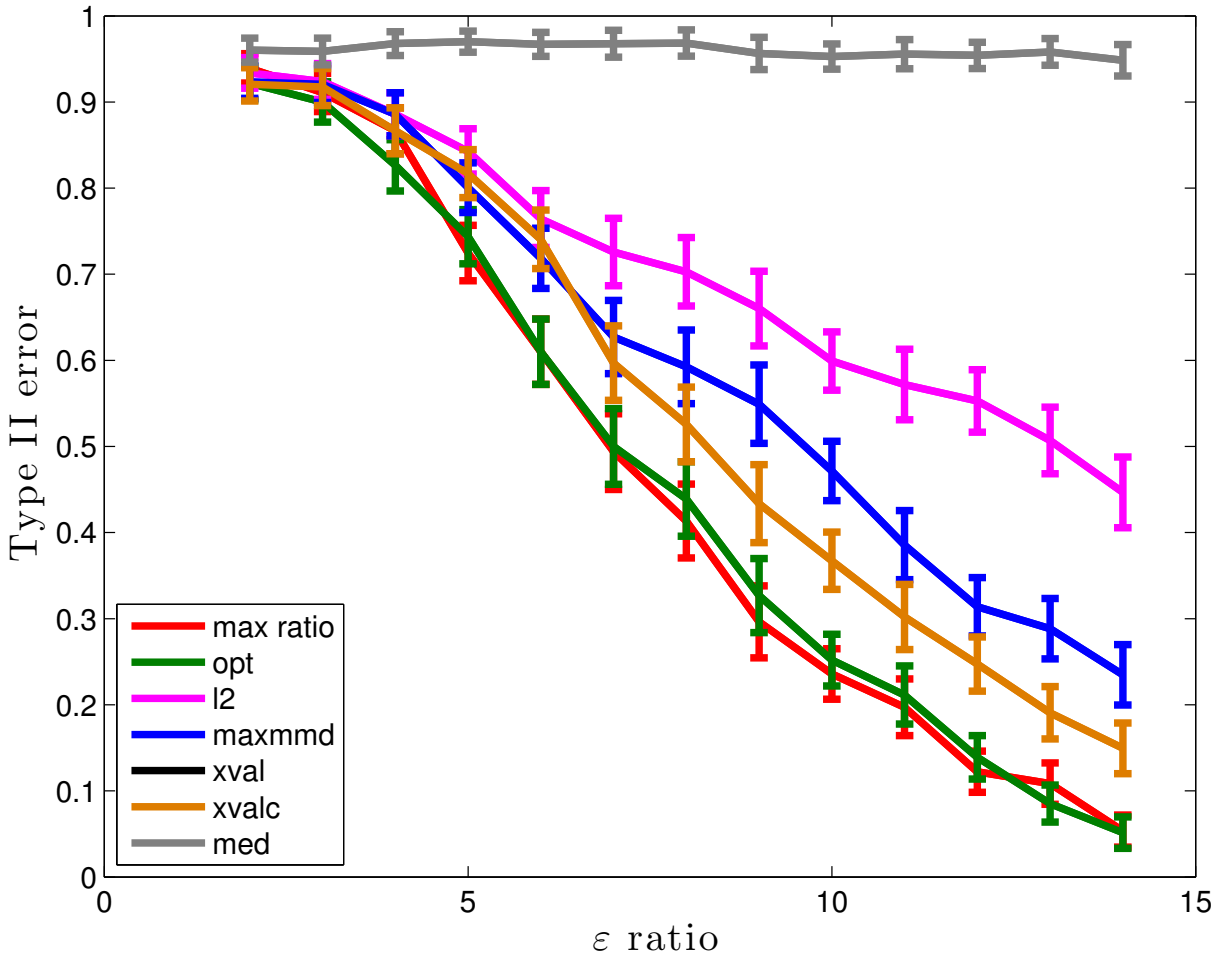
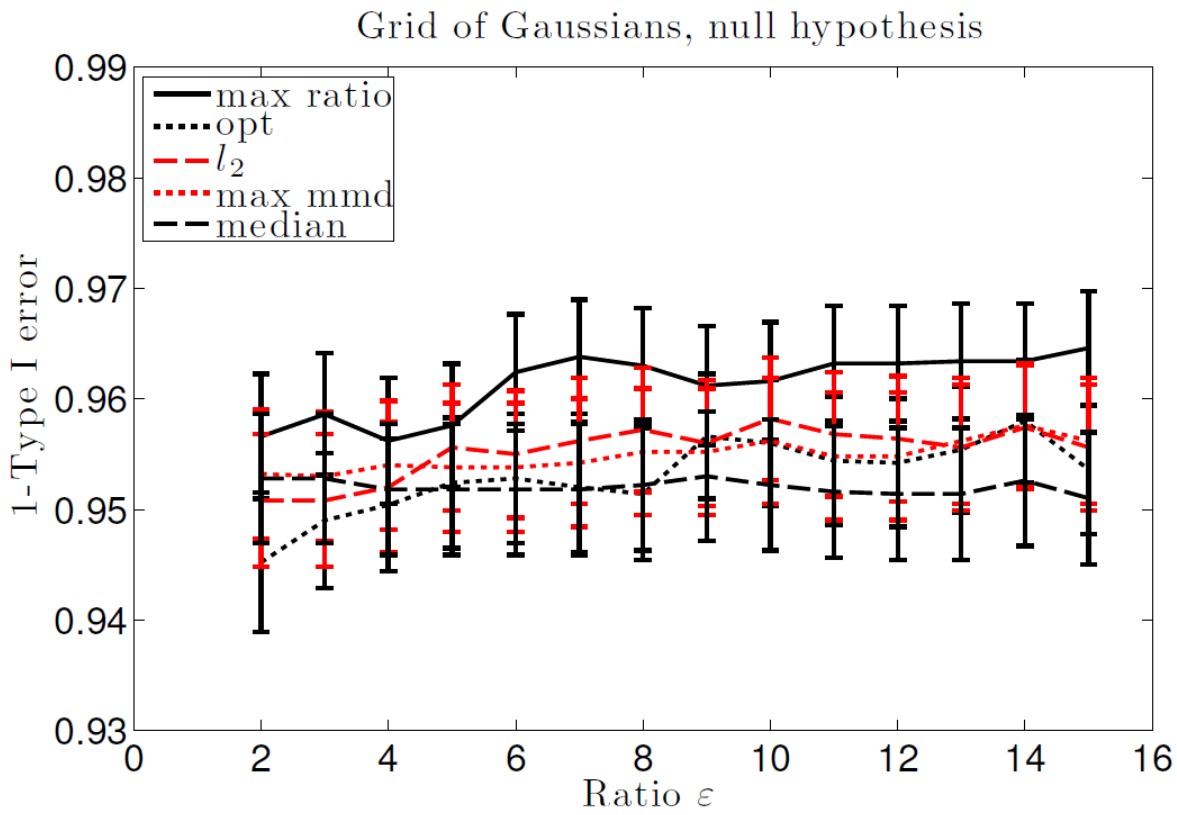


Figure: Parameters: $m = 10,000$ (for both training and testing)

Grid-mixtures: Type I error



Conclusions

- A criterion to explicitly optimize the (Hodges and Lehmann) asymptotic relative efficiency for the kernel two-sample test

Conclusions

- A criterion to explicitly optimize the (Hodges and Lehmann) asymptotic relative efficiency for the kernel two-sample test
- Consistency of a regularized empirical criterion, solved by a quadratic program

Conclusions

- A criterion to explicitly optimize the (Hodges and Lehmann) asymptotic relative efficiency for the kernel two-sample test
- Consistency of a regularized empirical criterion, solved by a quadratic program
- Both optimization and testing are performed with cost linear in the sample size (large-scale/streaming)

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

E-distance

- **Energy distance** (Székely, 1985; Székely and Rizzo 2004, 2005)

$$D_E(P, Q) = 2\mathbb{E}_{ZW} \|Z - W\|_2 - \mathbb{E}_{ZZ'} \|Z - Z'\|_2 - \mathbb{E}_{WW'} \|W - W'\|_2 \geq 0,$$

where $Z, Z' \overset{i.i.d.}{\sim} P$ and $W, W' \overset{i.i.d.}{\sim} Q$.

E-distance

- **Energy distance** (Székely, 1985; Székely and Rizzo 2004, 2005)

$$D_E(P, Q) = 2\mathbb{E}_{ZW} \|Z - W\|_2 - \mathbb{E}_{ZZ'} \|Z - Z'\|_2 - \mathbb{E}_{WW'} \|W - W'\|_2 \geq 0,$$

  where $Z, Z' \overset{i.i.d.}{\sim} P$ and $W, W' \overset{i.i.d.}{\sim} Q$.
- $D_E(P, Q) = 0$ if and only if $P = Q$.

Distance covariance (dCov)

Let X be a random vector on $\mathcal{X} = \mathbb{R}^p$ and Y a random vector on $\mathcal{Y} = \mathbb{R}^q$. The distance covariance $\mathcal{V}(X, Y)$ is defined via the norm of $f_{XY} - f_X f_Y$ in a weighted L_2 space on $\mathbb{R}^{p+q}$, i.e.,

$$\mathcal{V}^2(X, Y) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 \, w(t, s) \, dt \, ds,$$

for a particular choice of weight function given by:

$$w(t, s) = \frac{1}{c_p c_q} \cdot \frac{1}{\|t\|_2^{1+p} \, \|s\|_2^{1+q}},$$

where $c_d = \pi^{\frac{1+d}{2}} / \Gamma(\frac{1+d}{2})$, $d \geq 1$.

Distance covariance (dCov)

- **Distance covariance** (Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009; Lyons 2011)

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left\|X-X'\right\|_2\left\|Y-Y'\right\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_Y\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2 \\ &\quad - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2\right],\end{aligned}$$

where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$.

Distance covariance (dCov)

- **Distance covariance** (Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009; Lyons 2011)

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left\|X-X'\right\|_2\left\|Y-Y'\right\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_Y\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2 \\ &\quad - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2\right],\end{aligned}$$

  where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$.

- generalizes standard product-moment covariance (also leads to the notion of **distance correlation**)

Distance covariance (dCov)

- **Distance covariance** (Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009; Lyons 2011)

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left\|X-X'\right\|_2\left\|Y-Y'\right\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_Y\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2 \\ &\quad - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2\right],\end{aligned}$$

  where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$.
- generalizes standard product-moment covariance (also leads to the notion of **distance correlation**)
- $\mathcal{V}^2(X,Y)=0$ if and only if X and Y are independent

Characteristic function interpretation of MMD/HSIC

Let $k(z, z') = \kappa(z - z')$ be a translation invariant RKHS kernel on $\mathcal{Z}$, where $\kappa : \mathcal{Z} \to \mathbb{R}$ is a bounded continuous function. Using Bochner's theorem, κ is a Fourier transform of a **non-negative finite measure** Λ:

$$\kappa(\Delta) = \int e^{-\Delta^\top u} d\Lambda(u),$$

It follows (Gretton et al, 2009) that:

$$\gamma_k^2(P, Q) = \int_{\mathbb{R}^d} |f_Z(u) - f_W(u)|^2 \, d\Lambda(u).$$

Kernel approach = Energy approach?

Characteristic function interpretation

$$\gamma_k^2(P_{XY}, P_X P_Y) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2 \, d\Lambda(t,s)$$

$$\mathcal{V}^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{\|t\|_2^{1+p} \, \|s\|_2^{1+q}} \, dt \, ds$$

Characteristic function interpretation

$$\gamma_k^2(P_{XY}, P_X P_Y) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2 \, d\Lambda(t,s)$$

$$\mathcal{V}^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{\|t\|_2^{1+p} \, \|s\|_2^{1+q}} \, dt \, ds$$

just set $d\Lambda(t,s) = w(t,s)dtds$*?*

- $w(t,s)$ is not integrable, i.e.,$\kappa(\Delta) = \int \frac{e^{-\Delta^\top (t\ s)}}{\|t\|^{1+p} \|s\|^{1+q}} dtds$ does not converge, so there exist no **translation invariant** positive definite kernel that leads to distance covariance (Székely and Rizzo 2009, discussion by Gretton et al).

Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

Negative-type semimetric

Definition (Negative-type semimetric)

Let $\mathcal{Z}$ be a non-empty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$,

- $\rho(z, z') = 0$ if and only if $z = z'$, and $\rho(z, z') = \rho(z', z)$.

Then $(\mathcal{Z}, \rho)$ is said to be a semimetric space and ρ is called a semimetric on $\mathcal{Z}$. If, in addition, $\forall n \geq 2$, $z_1, \dots, z_n \in \mathcal{Z}$, and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$$

ρ is said to have negative type.

Negative-type semimetric

Definition (Negative-type semimetric)

Let $\mathcal{Z}$ be a non-empty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$,

- $\rho(z, z') = 0$ if and only if $z = z'$, and $\rho(z, z') = \rho(z', z)$.

Then $(\mathcal{Z}, \rho)$ is said to be a semimetric space and ρ is called a semimetric on $\mathcal{Z}$. If, in addition, $\forall n \geq 2$, $z_1, \dots, z_n \in \mathcal{Z}$, and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$$

ρ is said to have negative type.

- negative type $\Rightarrow D_{E,\rho}(P, Q) \geq 0$ (Lyons 2011)
- All Euclidean (and Hilbert) spaces are of negative type.

Distance-induced kernels

Distance-induced kernel: Let ρ be a semimetric on $\mathcal{Z}$ and $z_0 \in \mathcal{Z}$. Denote

$$k(z, z') = \frac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right]$$

Distance-induced kernels

Distance-induced kernel: Let ρ be a semimetric on $\mathcal{Z}$ and $z_0 \in \mathcal{Z}$. Denote

$$k(z, z') = \frac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right]$$

translation variant!

Distance-induced kernels

Distance-induced kernel: Let ρ be a semimetric on $\mathcal{Z}$ and $z_0 \in \mathcal{Z}$. Denote

$$k(z, z') = \frac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right]$$

translation variant!

Proposition

*k is a valid (psd) kernel **if and only if** ρ is of negative type. Conversely, if k is a psd kernel, then:*

$$\rho(z, z') = k(z, z) + k(z', z') - 2k(z, z')$$

is a negative-type semimetric (generated by k).

Main results

Theorem

Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type and let k be any kernel that generates ρ. Then,

$$D_{E,\rho}(P, Q) = 2\gamma_k^2(P, Q), \qquad \forall P, Q \in \mathcal{M}_k^1(\mathcal{Z}).$$

Main results

Theorem

Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type and let k be any kernel that generates ρ. Then,

$$D_{E,\rho}(P, Q) = 2\gamma_k^2(P, Q), \qquad \forall P, Q \in \mathcal{M}_k^1(\mathcal{Z}).$$

Theorem

Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semimetric spaces of negative type, and let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be any two kernels on $\mathcal{X}$ and $\mathcal{Y}$ that generate $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$, respectively. Then, if $(X, Y) \sim P_{XY}$, with marginals $P_X \in \mathcal{M}_{k_{\mathcal{X}}}^2(\mathcal{X})$, $P_Y \in \mathcal{M}_{k_{\mathcal{Y}}}^2(\mathcal{Y})$,

$$\mathcal{V}^2_{\rho_{\mathcal{X}}, \rho_{\mathcal{Y}}}(X, Y) = 4HSIC^2(X, Y; k_{\mathcal{X}}, k_{\mathcal{Y}}).$$

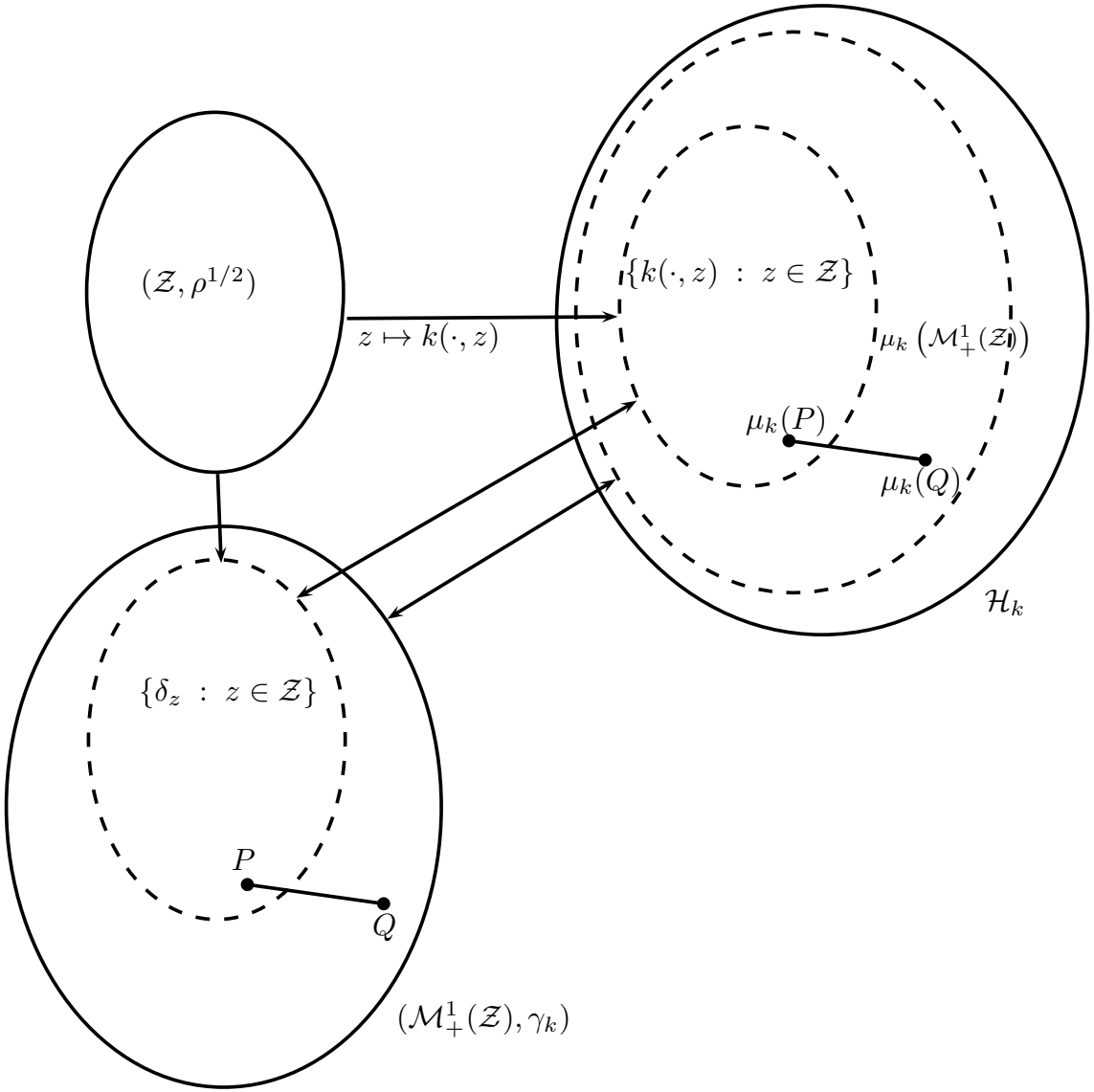
Main results (2)

- In testing, one simply replaces population expressions with their empirical versions (energy distance $D_{E,\rho}(\hat{P}, \hat{Q})$ and MMD $\gamma_k(\hat{P}, \hat{Q})$ between empirical distributions)

Main results (2)

- In testing, one simply replaces population expressions with their empirical versions (energy distance $D_{E,\rho}(\hat{P}, \hat{Q})$ and MMD $\gamma_k(\hat{P}, \hat{Q})$ between empirical distributions)
- Thus, kernel-based and energy-based statistics are equivalent under the above moment-assumptions

Two ways to induce a metric on $\mathcal{M}_+^1(\mathcal{Z})$



Outline

1. Introduction and Motivation
2. RKHS/kernel embedding/MMD
3. Estimating MMD / Testing
4. Kernel selection in large-scale two-sample tests
   - Asymptotic efficiency criterion
   - Experiments
5. Equivalence to energy distance/distance covariance
   - Beyond Euclidean metrics
   - Spectral testing for e-distance

Spectral test

- (Gretton et al, 2009)

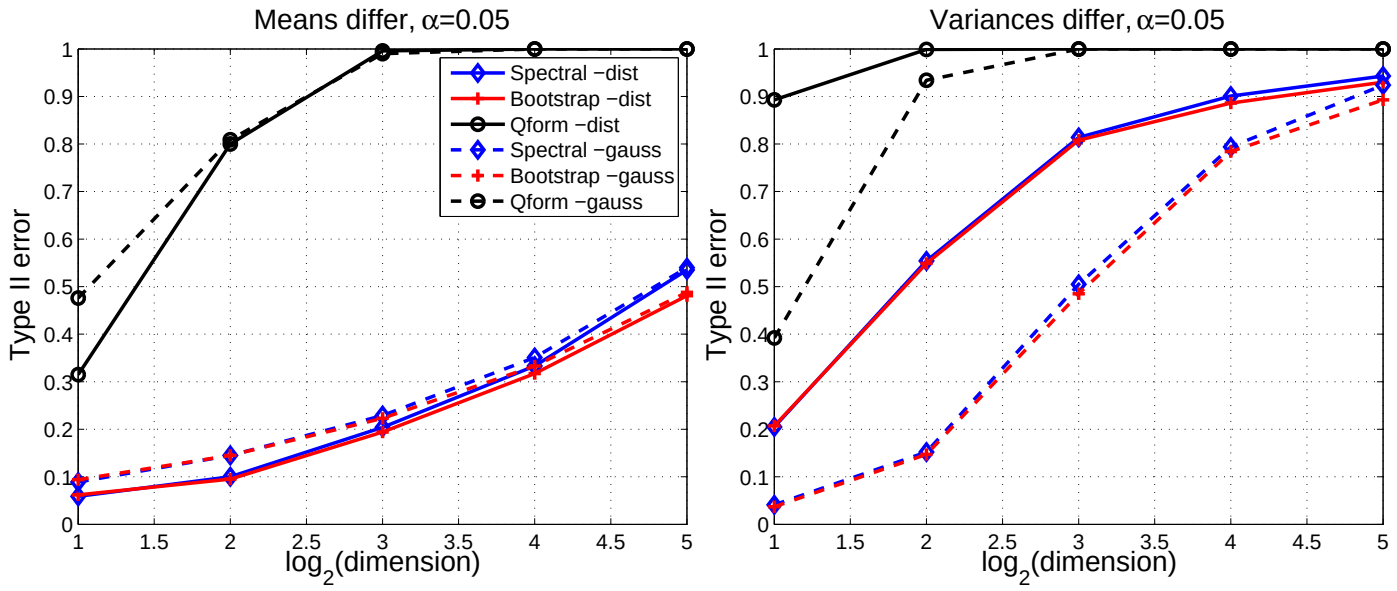
$$\frac{m}{2}\hat{\eta}_{k,V}(\mathbf{z},\mathbf{w}) \quad \rightsquigarrow \quad \sum_{i=1}^{\infty}\lambda_i N_i^2.$$

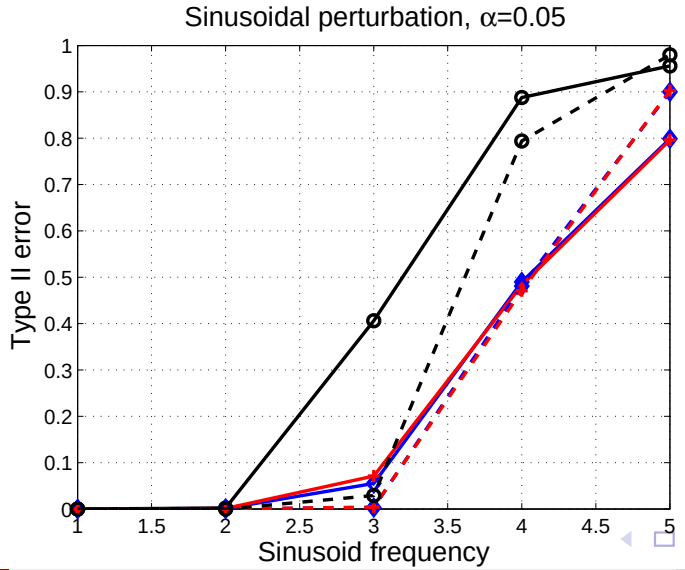
- Compute the Gram matrix K on the aggregated samples, $K_{ij} = k(u_i, u_j)$, $\mathbf{u} = [\mathbf{z}\ \mathbf{w}]$
- Compute the spectrum of its centred version $\tilde{K} = HKH$ (surrogate for $S_{\tilde{k}_P}$)
- cost: $\mathcal{O}(m^3)$ instead of $\mathcal{O}(m^4)$ for the permutation test.

Quadratic form test

- (Szekely et al, 2007) also express distance-statistics as a quadratic form Q of centered Gaussian random variables (no method to estimate coefficients is given)
- Test based on $\mathbb{P}\left\{Q \geq \left(\Phi^{-1}(1-\alpha/2)^2\right)\right\} \leq \alpha$, valid for $0 < \alpha \leq 0.215$, valid for all quadratic forms Q, with $\mathbb{E}Q = 1$
- When applied to the dCov statistic, the upper bound of α is achieved if X and Y are **independent Bernoulli** - over-conservative in general

Two-sample testing results





Conclusions

- Distance-based statistics of Szekely et al are a special case of the RKHS framework.

Conclusions

- Distance-based statistics of Szekely et al are a special case of the RKHS framework.
- Conversely, RKHS-based statistics have a clear interpretation in terms of implicitly imposing a (semi)metric onto the original space.

Conclusions

- Distance-based statistics of Szekely et al are a special case of the RKHS framework.
- Conversely, RKHS-based statistics have a clear interpretation in terms of implicitly imposing a (semi)metric onto the original space.
- New way to estimate the null distribution of distance-statistics through the link with kernels.

Conclusions

- Distance-based statistics of Szekely et al are a special case of the RKHS framework.
- Conversely, RKHS-based statistics have a clear interpretation in terms of implicitly imposing a (semi)metric onto the original space.
- New way to estimate the null distribution of distance-statistics through the link with kernels.
- For problem settings defined most naturally in terms of some given distances, and where these distances are of negative type, RKHS machinery can be brought to bear (ISOMAP $\leftrightarrow$ Kernel PCA).

References

- A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, in *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.
- D. Sejdinovic, A. Gretton, B. Sriperumbudur and K. Fukumizu, **Hypothesis testing using pairwise distances and associated kernels**, in *Proc. International Conference on Machine Learning ICML*, 2012.
- D. Sejdinovic, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**, in review [arXiv:1207.6076].
- A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf and A. Smola, **A Kernel Two-Sample Test**, *Journal of Machine Learning Research*, 13:723-773, 2012.
- G. Székely and M. Rizzo, **Brownian distance covariance** (with discussion). *Ann. Appl. Stat.*, 4(3):1233–1303, 2009.
- R. Lyons, **Distance covariance in metric spaces**. *Ann. Probab.* (to appear)

Test procedure

The testing procedure is as follows:

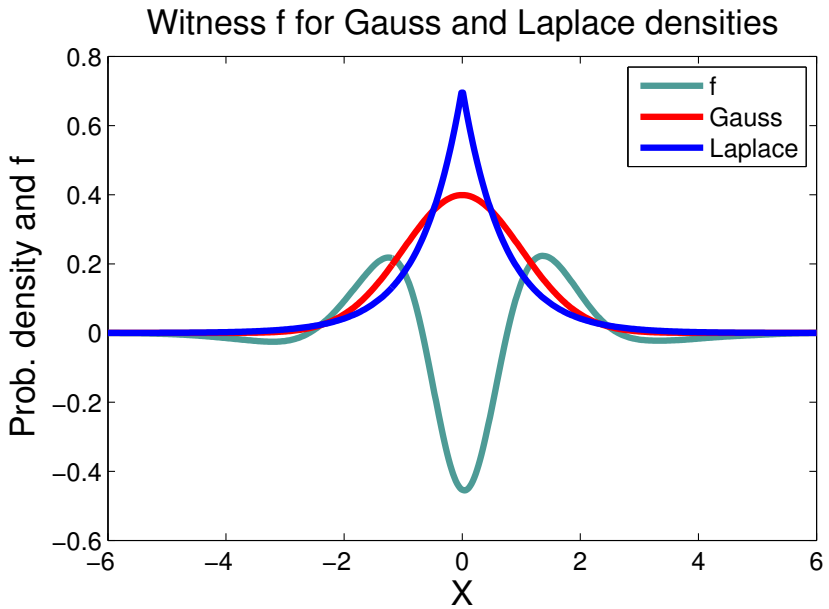
1. Split the data into training and testing
2. On the training data:
   1. Compute $\hat{\eta}_u$ for all $k_u \in \mathcal{K}$
   2. If *at least one* $\hat{\eta}_u > 0$, compute $\hat{Q}$, and solve the QP to get $\hat{\beta}^*$, else choose a single k_u that maximizes $\hat{\eta}_u / \hat{\sigma}_{u,\lambda}$
3. On the test data:
   1. Compute $\check{\eta}_{\hat{k}_*}$ using $\hat{k}_* = \sum_{u=1}^{d} \hat{\beta}_u^* k_u$
   2. Compute test threshold $\check{t}_{\alpha,\hat{k}_*}$ using $\check{\sigma}_{\hat{k}_*}$
4. Reject null if $\check{\eta}_{\hat{k}_*} > \check{t}_{\alpha,\hat{k}_*}$

MMD as integral probability metric

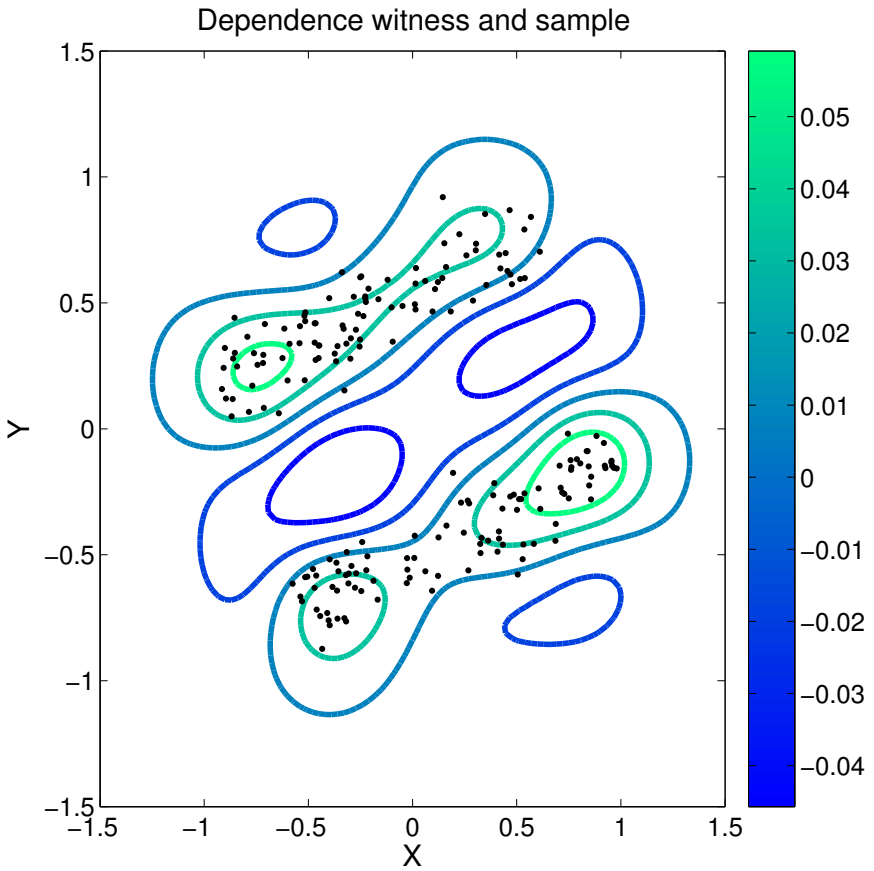
- An alternative interpretation of MMD is as an integral probability metric (Müller, 1997), i.e.,

$$\gamma_k(P, Q) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \left[\mathbb{E}_{Z \sim P} f(Z) - \mathbb{E}_{W \sim Q} f(W)\right].$$

- Supremum acheived at the "witness function"
$f = (\mu_k(P) - \mu_k(Q)) / \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$.



HSIC as integral probability metric



- $HSIC^2(X, Y; k_{\mathcal{X}}, k_{\mathcal{Y}}) = \|\mu_k(P_{XY}) - \mu_k(P_X P_Y)\|^2_{\mathcal{H}_k}$
- witness lies in $\mathcal{H}_k$, the RKHS of functions on $\mathcal{X} \times \mathcal{Y}$