STAT G8325
Gaussian Processes and Kernel Methods
§10: Kernel Statistical Tests

John P. Cunningham

Department of Statistics
Columbia University

Outline

Outline

Administrative interlude

Following up from last time: mean embedding

Hilbert-Schmidt operators

Context: kernel statistical tests

Hilbert-Schmidt independence criterion [GBSS05]

Kernel two-sample tests [GBR+12]

References

Progress...

Week	Lectures	Content
9	Nov 16, 18	Introduction to kernel methods • [SSM98] (intentionally light reading; work on projects)
10	Nov 23, Dec 2	Kernel statistical tests • [GBR$^+$12] (intentionally light reading; work on projects)
11	Nov 30	(Project progress report)

- HW4 due this past weekend.
- HW5 due next Monday:
  - Present 2-3 minutes of your project progress, in class.
  - Solicit feedback from all of us.
  - Identify key issues.
  - Attendance here is important.
- Class will not be held this Wednesday Nov 25.
- We will contextualize kernel methods today.

Outline

Administrative interlude

Following up from last time: mean embedding

Hilbert-Schmidt operators

Context: kernel statistical tests

Hilbert-Schmidt independence criterion [GBSS05]

Kernel two-sample tests [GBR+12]

References

Mean embeddings

- In §09, we discussed $\mu_P = E_{P(x)}(\phi(x)) \in \mathcal{H}$.
- Note $Ef \triangleq E_{P(x)}(f(x))$ is a linear operator $\mathcal{H} \to \mathbb{R}$.
- Thus E bounded $\Rightarrow \mu_P \in \mathcal{H}$.
  - Due to Riesz: $Ef = E_{P(x)}(f(x)) = \langle f, \mu_P \rangle_{\mathcal{H}}$.
  - Take a moment to make sense of that statement.
- Then:

$$
\begin{aligned}
|Ef| &= |E_{P(x)}(f(x))| \\
&\leq E_{P(x)}(|f(x)|) \\
&= E_{P(x)}(|\langle f, \phi(x) \rangle_{\mathcal{H}}|) \\
&\leq E_{P(x)}\left(\sqrt{k(x,x)}\right) \|f\|_{\mathcal{H}}.
\end{aligned}
$$

- Thus $\mu_P \in \mathcal{H}$ exists if $E_{P(x)}\left(\sqrt{k(x,x)}\right) < \infty$.

Outline

Products of kernels are kernels

- Back in §04, we hinted that $k = k^1 k^2$ is a kernel for kernels k^1, k^2.
- We hinted at this fact via Euclidean feature maps $\phi^1 : \mathcal{X} \to \mathbb{R}^{d_1}$:

$$
\begin{aligned}
k(x,x') &= k^1(x,x')k^2(x,x') \\
&= \phi^1(x)^\top \phi^1(x')\phi^2(x)^\top \phi^2(x') \\
&= tr\left(\phi^1(x')^\top \phi^1(x)\phi^2(x)^\top \phi^2(x')\right) \\
&= tr\left(\phi^2(x')\phi^1(x')^\top \phi^1(x)\phi^2(x)^\top\right) \\
&= \left(\phi^1(x')\phi^2(x')^\top\right)^\top \left(\phi^1(x)\phi^2(x)^\top\right) \\
&= \left\langle \left(\phi^1(x')\phi^2(x')^\top\right), \left(\phi^1(x)\phi^2(x)^\top\right)\right\rangle_{\mathbb{R}^{d_1 \times d_2}},
\end{aligned}
$$

  thus showing that the product is itself a kernel.

- We will derive this in more generality so as to motivate Hilbert-Schmidt operators (and to satisfy ourselves that this property is true in generality).

Product of kernels are kernels

- Consider $f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2$ for two separable rkhs with rk k^1, k^2.
- For any $f \in \mathcal{H}_2$, define the tensor product operator $f_1 \otimes f_2 : \mathcal{H}_2 \to \mathcal{H}_1$ as:

$$(f_1 \otimes f_2)f = \langle f_2, f \rangle_{\mathcal{H}_2} f_1.$$

- The above form reminds us of:
  - typical Euclidean outer products: $(ab^\top)c = \langle b, c \rangle a$.
  - the Kronecker trick: $(A \otimes B^\top)vec(C) = vec\left((B^\top C)A^\top\right)$.
- More generally, $f_1 \otimes f_2$ is just a (rank one) operator $L : \mathcal{H}_2 \to \mathcal{H}_1$.
- When such an operator (not necessarily rank one) is:
  - bounded: $\|L\| = \sup_f \frac{\|Lf\|_{\mathcal{H}_1}}{\|f\|_{\mathcal{H}_2}} < \infty$.
  - has finite *Hilbert-Schmidt norm* (for an onb $\{\varphi_i^2\}_i$ of $\mathcal{H}_2$):

$$\|L\|_{HS}^2 = \sum_{i \in \mathbb{N}} \|L\varphi_i^2\|_{\mathcal{H}_1}^2,$$

  ...then L is called a *Hilbert-Schmidt operator*. We say $L \in HS(\mathcal{H}_2, \mathcal{H}_1)$.
- $HS(\mathcal{H}_2, \mathcal{H}_1)$ is itself a Hilbert space, using the implied inner product

$$\langle L, M \rangle_{HS} = \sum_{i \in \mathbb{N}} \left\langle L\varphi_i^2, M\varphi_i^2 \right\rangle_{\mathcal{H}_1}.$$

Product of kernels are kernels

- Show $f_1 \otimes f_2 \in HS(\mathcal{H}_2, \mathcal{H}_1)$:
  - The operator $f_1 \otimes f_2$ is bounded:

$$\begin{aligned}\|f_1 \otimes f_2\| &= \sup_f \frac{\|(f_1 \otimes f_2) f\|_{\mathcal{H}_1}}{\|f\|_{\mathcal{H}_2}} \\ &= \sup_f \frac{\| \langle f_2, f \rangle_{\mathcal{H}_2} f_1 \|_{\mathcal{H}_1}}{\|f\|_{\mathcal{H}_2}} \\ &= \sup_f \frac{| \langle f_2, f \rangle_{\mathcal{H}_2} | \|f_1\|_{\mathcal{H}_1}}{\|f\|_{\mathcal{H}_2}} \\ &= \|f_2\|_{\mathcal{H}_2} \|f_1\|_{\mathcal{H}_1}.\end{aligned}$$

  - The operator $f_1 \otimes f_2$ has finite HS norm:

$$\begin{aligned}\|f_1 \otimes f_2\|_{HS}^2 &= \sum_{i \in \mathbb{N}} \|(f_1 \otimes f_2) \varphi_i^2\|_{\mathcal{H}_1}^2 \\ &= \sum_{i \in \mathbb{N}} \| \left\langle f_2, \varphi_i^2 \right\rangle_{\mathcal{H}_2} f_1 \|_{\mathcal{H}_1}^2 \\ &= \|f_1\|_{\mathcal{H}_1}^2 \sum_{i \in \mathbb{N}} \left\langle f_2, \varphi_i^2 \right\rangle_{\mathcal{H}_2} \\ &= \|f_1\|_{\mathcal{H}_1}^2 \|f_2\|_{\mathcal{H}_1}^2.\end{aligned}$$

  - Thus $f_1 \otimes f_2 \in HS(\mathcal{H}_2, \mathcal{H}_1)$.

Product of kernels

- Now consider

$$\begin{aligned}\langle \phi \otimes \phi', L \rangle_{HS} &= \sum_{i \in \mathbb{N}} \langle (\phi \otimes \phi')\, \varphi_i, L\varphi_i \rangle_{\mathcal{H}_1} \\ &= \sum_{i \in \mathbb{N}} \langle \phi', \varphi_i \rangle_{\mathcal{H}_2} \langle \phi, L\varphi_i \rangle_{\mathcal{H}_1} \\ &= \left\langle \phi, \sum_{i \in \mathbb{N}} \langle \phi', \varphi_i \rangle_{\mathcal{H}_2} L\varphi_i \right\rangle_{\mathcal{H}_1} \\ &= \left\langle \phi, L \sum_{i \in \mathbb{N}} \langle \phi', \varphi_i \rangle_{\mathcal{H}_2} \varphi_i \right\rangle_{\mathcal{H}_1} \\ &= \langle \phi, L\phi' \rangle_{\mathcal{H}_1}.\end{aligned}$$

- In particular, if $L = \phi_2 \otimes \phi_2' \in HS(\mathcal{H}_2, \mathcal{H}_1)$, then:

$$\begin{aligned}\langle \phi_1 \otimes \phi_1', \phi_2 \otimes \phi_2' \rangle_{HS} &= \langle \phi_1, (\phi_2 \otimes \phi_2')\phi_1' \rangle_{\mathcal{H}_1} \\ &= \left\langle \phi_1, \langle \phi_1', \phi_2' \rangle_{\mathcal{H}_2} \phi_2 \right\rangle_{\mathcal{H}_1} \\ &= \langle \phi_1, \phi_2 \rangle_{\mathcal{H}_1} \langle \phi_1', \phi_2' \rangle_{\mathcal{H}_2}.\end{aligned}$$

- Now finally, the product $k = k^1 k^2$ is a kernel:

$$k(x, x') = k^1(x, x')k^2(x, x') = \langle \phi_1, \phi_1' \rangle_{\mathcal{H}_1} \langle \phi_2, \phi_2' \rangle_{\mathcal{H}_2} = \langle \phi_1 \otimes \phi_1', \phi_2 \otimes \phi_2' \rangle_{HS}.$$

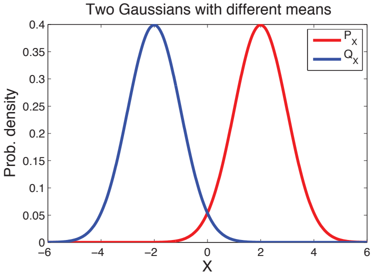
Outline

Context

- We have covered the kernel versions of a number of hugely important methods:
  - Kernel ridge regression
  - Kernel mean estimation
  - Kernel PCA
  - (Kernel SVM, nearest neighbors, k-means, ...)
- The broader context: we can answer more general questions by kernelizing our canon of statistical methods on linear features.
- However, I do acknowledge that sometimes 'kernelized' methods seem cute as opposed to fundamentally interesting.
- Hereafter we will focus on some statistical testing applications that, in my view (and that of many others) are of quite fundamental importance...

Statistical independence tests

- Detecting statistical independence between data sets is a fundamental (and hugely important) problem in statistics.
- Classic: the two-sample t-test for normals with different means.

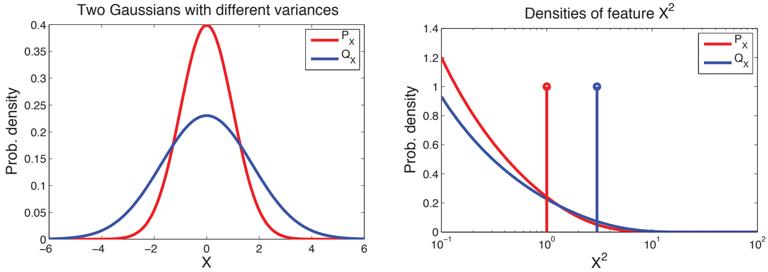


- How do we think about statistical independence more generally?
- Certainly comparing p_{xy} vs $p_x p_y$ for rv's X, Y is the ideal step, but approaching this with finite data is usually quite challenging.

Note for remainder: we're covering Gretton's papers, so we'll often use his figures.

Statistical independence tests

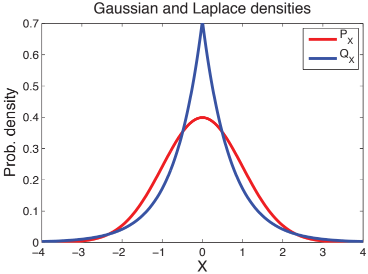
- What if we have different features (other than the mean) that define the difference between two distributions?
- Example: normals with different variance.
- Consider a feature map $\phi(x) = x^2$.



- If these distributions have means 'far enough apart', then we can conclude two distributions with independent variance.
- Can we now see where this all is going?

Statistical independence tests

- What if we have two distributions with same means and variance but different higher order features?



- We can use an rkhs to give us an infinite feature map.
- We can then compute the means of each of these features.
- And we can perform a statistical test to see if two samples of data have the same mean (of a bunch of features) or different.
- This basic strategy underlies much literature; we will highlight two particularly excellent examples.

Outline

Administrative interlude

Following up from last time: mean embedding

Hilbert-Schmidt operators

Context: kernel statistical tests

Hilbert-Schmidt independence criterion [GBSS05]

Kernel two-sample tests [GBR+12]

References

Setup

- Some basic assumptions for what follows:
  - Assume a distribution P_x on $\mathcal{X}$ (with Borel σ-algebra).
  - Assume a distribution P_y on $\mathcal{Y}$ (with Borel σ-algebra).
  - Assume representers and rkhs $\phi_x : \mathcal{X} \to \mathcal{H}_x$ and $\phi_y : \mathcal{Y} \to \mathcal{H}_y$.
  - All rkhs are separable (easiest sufficient: continuous k on separable $\mathcal{X}$)

- As before we define the mean: $E_x(f(x)) = E_x\left(\langle \phi_x, f \rangle_{\mathcal{H}_x}\right) \triangleq \langle \mu_x, f \rangle_{\mathcal{H}_x}$.

- New: define the *cross-covariance* operator:

$$\begin{aligned} C_{xy} &\triangleq E_{xy}\left((\phi_x - \mu_x) \otimes (\phi_y - \mu_y)\right) \\ &= E_{xy}\left(\phi_x \otimes \phi_y\right) - \mu_x \otimes \mu_y \\ &\triangleq \tilde{C}_{xy} - M_{xy}. \end{aligned}$$

  ...recall our focus on tensor products and Hilbert-Schmidt earlier today.

- Critically, note that $C_{xy} \in HS(\mathcal{H}_x, \mathcal{H}_y)$.

Hilbert-Schmidt independence criteria

- [GBSS05] defines the *Hilbert-Schmidt Independence Criterion* as:

$$HSIC\left(p_{xy}, \mathcal{H}_x, \mathcal{H}_y\right) \triangleq \|C_{xy}\|_{HS}.$$

- In concrete 'kernel' terms:

$$\begin{aligned}
\|C_{xy}\|_{HS}^2 &= \left\langle E_{xy}\left(\phi_x \otimes \phi_y\right) - \mu_x \otimes \mu_y \; , \; E_{x'y'}\left(\phi_{x'} \otimes \phi_{y'}\right) - \mu_{x'} \otimes \mu_{y'} \right\rangle_{HS} \\
&= E_{xyx'y'}\left(\left\langle \phi_x \otimes \phi_y, \phi_{x'} \otimes \phi_{y'} \right\rangle_{HS}\right) - 2E_{xy}\left(\langle \mu_x \otimes \mu_y, \phi_x \otimes \phi_y \rangle_{HS}\right) \\
&\quad + \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{HS}. \\
&= E_{xyx'y'}\left(k_x(x,x')k_y(y,y')\right) - 2E_{xy}\left(E_{x'}\left(k_x(x,x')\right) E_{y'}\left(k_y(y,y')\right)\right) \\
&\quad + E_{xx'}\left(k_x(x,x')\right) E_{yy'}\left(k_y(y,y')\right).
\end{aligned}$$

Notice that we have used our product kernel knowledge in a nontrivial way!

- We must now:
  - Compute this quantity in a finite setting
  - Show how we can use it for an independence test.

Finite HSIC

- Say we have a finite data set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, drawn from the joint distribution p_{xy}.
- The following estimator is proposed:

$$
\begin{aligned}
HSIC(D, \mathcal{H}_x, \mathcal{H}_y) &\triangleq \frac{1}{(n-1)^2} tr\left(K_x(I - \frac{1}{n}11^\top)K_y(I - \frac{1}{n}11^\top)\right) \\
&= \frac{1}{(n-1)^2} tr\left(K_x K_y - 2\frac{1}{n}11^\top K_x K_y + \frac{1}{n}11^\top K_x \frac{1}{n}11^\top K_y\right).
\end{aligned}
$$

where K_x and K_y are the appropriate kernel matrices.

- Interpret each use of $\frac{1}{n}11^\top$ as a mean operation μ_x... a sensible estimator.
- (This estimator is also biased, and will be improved in subsequent work.)
- [GBSS05] to use large deviation bounds to show this finite expectation is appropriately behaved.

Independence testing using HSIC

- Important theorem: $\|C_{xy}\|_{HS} = 0 \Leftrightarrow x, y$ are independent.
- Accordingly, set an indicator function:

  $$\mathbb{1}\left(HSIC(D, \mathcal{H}_x, \mathcal{H}_y) > \gamma_\alpha\right),$$

  where γ_α is a suitably chosen constant such that this rejection test will have miss rate less than α. This is the typical setup for a rejection test.
- Based on the bound we skipped above, this value has form $C\sqrt{-\log \alpha / n}$.
- And no particular clarity is given on how to choose C.

Results

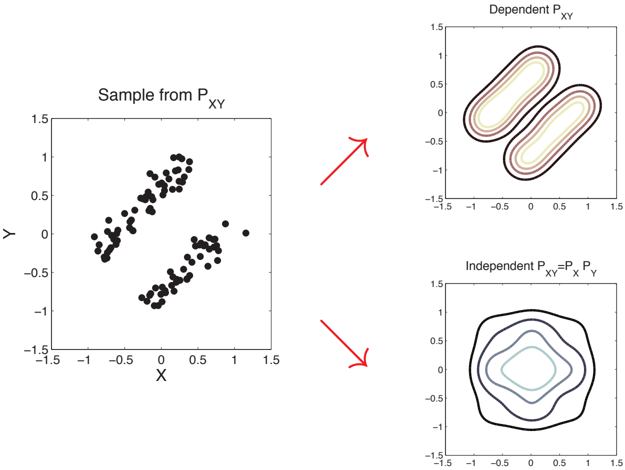
n	m	Rep.	FICA	Jade	IMAX	RAD	CFIC	KCC	COg	COl	KGV	KMIg	KMIl	HSg	HSl
2	250	1000	10.5 ± 0.4	9.5 ± 0.4	44.4 ± 0.9	**5.4 ± 0.2**	7.2 ± 0.3	7.0 ± 0.3	7.8 ± 0.3	7.0 ± 0.3	**5.3 ± 0.2**	6.0 ± 0.2	5.7 ± 0.2	5.9 ± 0.2	5.8 ± 0.3
2	1000	1000	6.0 ± 0.3	5.1 ± 0.2	11.3 ± 0.6	**2.4 ± 0.1**	3.2 ± 0.1	3.3 ± 0.1	3.5 ± 0.1	2.9 ± 0.1	**2.3 ± 0.1**	2.6 ± 0.1	**2.3 ± 0.1**	2.6 ± 0.1	**2.4 ± 0.1**
4	1000	100	5.7 ± 0.4	5.6 ± 0.4	13.3 ± 1.1	**2.5 ± 0.1**	3.3 ± 0.2	4.5 ± 0.4	4.2 ± 0.3	4.6 ± 0.6	3.1 ± 0.6	4.0 ± 0.7	3.5 ± 0.7	2.7 ± 0.1	**2.5 ± 0.2**
4	4000	100	3.1 ± 0.2	2.3 ± 0.1	5.9 ± 0.7	**1.3 ± 0.1**	1.5 ± 0.1	2.4 ± 0.5	1.9 ± 0.1	1.6 ± 0.1	1.4 ± 0.1	1.4 ± 0.05	**1.2 ± 0.05**	**1.3 ± 0.05**	**1.2 ± 0.05**
8	2000	50	4.1 ± 0.2	3.6 ± 0.2	9.3 ± 0.9	**1.8 ± 0.1**	2.4 ± 0.1	4.8 ± 0.9	3.7 ± 0.9	5.2 ± 1.3	2.6 ± 0.3	2.1 ± 0.1	**1.9 ± 0.1**	**1.9 ± 0.1**	**1.8 ± 0.1**
8	4000	50	3.2 ± 0.2	2.7 ± 0.1	6.4 ± 0.9	**1.3 ± 0.05**	1.6 ± 0.1	2.1 ± 0.2	2.0 ± 0.1	1.9 ± 0.1	1.7 ± 0.2	**1.4 ± 0.1**	**1.3 ± 0.05**	**1.4 ± 0.05**	**1.3 ± 0.05**
16	5000	25	2.9 ± 0.1	3.1 ± 0.3	9.4 ± 1.1	**1.2 ± 0.05**	1.7 ± 0.1	3.7 ± 0.6	2.4 ± 0.1	2.6 ± 0.2	1.7 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	**1.3 ± 0.05**	**1.3 ± 0.05**

- Benchmark used is demixing data via ICA.
- HSIC (and others) are used to test whether ICA has recovered true independent components.
- Sample size m, repetitions rep, dimensionality n, measure is Amari divergence (to quantify independence of resulting distributions).
- Takeaway: $HSIC$ (in the gaussian g and laplace l kernel cases) is performant with many bespoke algorithms for ICA.

Outline

Dependence tests

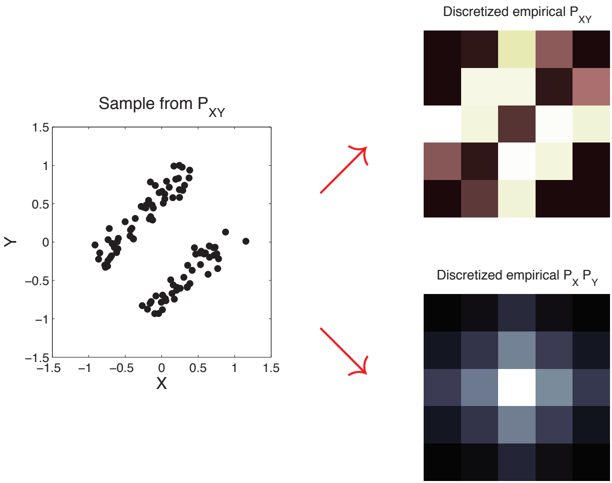
- Detecting statistical independence between variables is a fundamental (and hugely important) problem in statistics.



(from Gretton's recent MLSS)

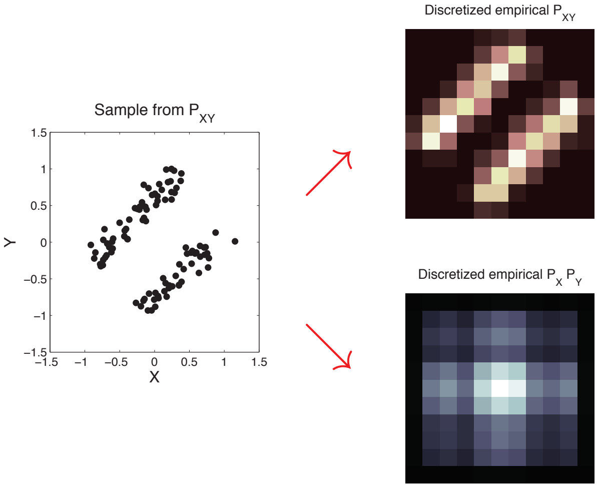
Dependence tests

- Detecting statistical independence between variables is a fundamental (and hugely important) problem in statistics.

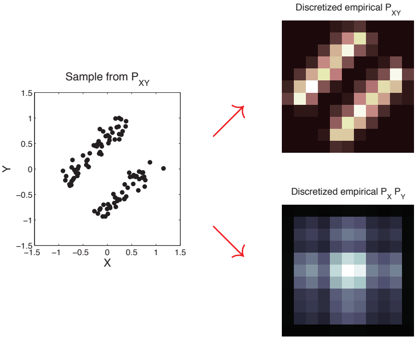


Dependence tests

- Detecting statistical independence between variables is a fundamental (and hugely important) problem in statistics.



Dependence tests



- Discretize and use a statistical test for categorical variables?
- Unfortunately, the curse of dimensionality quickly leads to failure.
- Intuition: each bin needs adequate data to distinguish $p_x p_y$ from p_{xy}.

Revisiting Hilbert-Schmidt independence criteria

- Earlier we defined the *Hilbert-Schmidt Independence Criterion* as:

$$\begin{aligned} HSIC^2\left(p_{xy}, \mathcal{H}_x, \mathcal{H}_y\right) &\triangleq \|C_{xy}\|^2_{HS} \\ &= \|E_{xy}(\phi_x \otimes \phi_y) - \mu_x \otimes \mu_y\|^2_{HS} \\ &\triangleq \|\mu_{p_{xy}} - \mu_{p_x p_y}\|^2_{HS}, \end{aligned}$$

  where the last line defines the mean (HS) operators of the joint and the product of the marginals.

- This shows us that HSIC is a distance between mean feature maps.

- Also we proved conditions for existence of μ_p (...$E_p\sqrt{k(x,x)} < \infty$).

- This reminds us that testing dependence between two rv's X and Y really boils down to some distance measure between the joint and the product of their marginals.

Maximum mean discrepancy [GBR+12]

- Again, our fundamental problem of interest is:
  - Given $x_i \sim_{iid} p$ and $y_j \sim_{iid} q$
  - Is $p \neq q$?
  - Certainly $p_{xy} \neq p_x p_y$ is such an example.

Apologies: x and y now have different roles.

- Define the *maximum mean discrepancy* as:

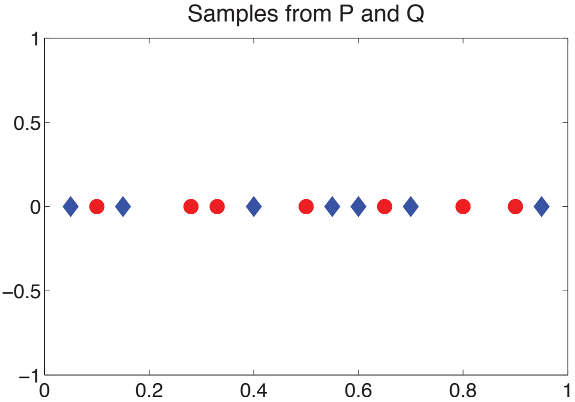
$$MMD(\mathcal{H}, p, q) \triangleq \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y))\right).$$

- Break this down:
  - Take a set of smooth functions $\mathcal{H}$...
  - ...(which of course is going to be something like an rkhs)...
  - And calculate the difference in the expectation of that function under p and q.
  - Think this should be small when $p = q$, big when $p \neq q$.

Maximum mean discrepancy

- Maximum mean discrepancy:

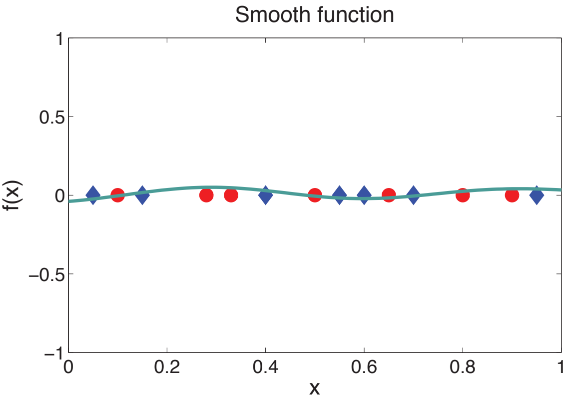
$$MMD(\mathcal{H}, p, q) \triangleq \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y)) \right).$$



Maximum mean discrepancy

- Maximum mean discrepancy:

$$MMD(\mathcal{H}, p, q) \triangleq \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y)) \right).$$

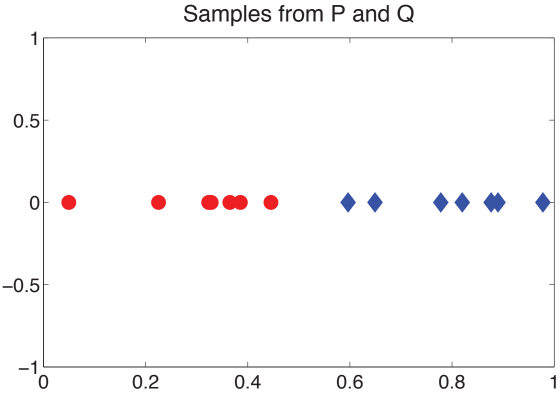


- Small MMD!

Maximum mean discrepancy

- Maximum mean discrepancy:

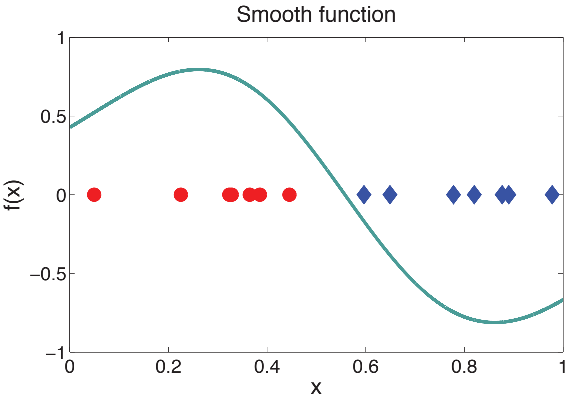
$$MMD(\mathcal{H}, p, q) \triangleq \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y)) \right).$$



Maximum mean discrepancy

- Maximum mean discrepancy:

$$MMD(\mathcal{H}, p, q) \triangleq \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y)) \right).$$



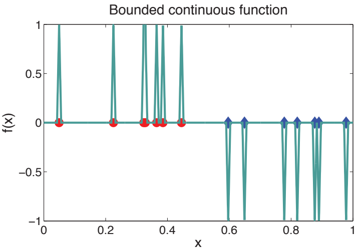
- Large MMD!

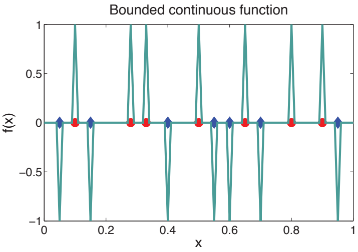
Maximum mean discrepancy

- Maximum mean discrepancy:

$$MMD(\mathcal{H}, p, q) \triangleq \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y)) \right).$$

- Smoothness matters!





Maximum mean discrepancy

- Assume μ_p, μ_q exist as before. Notice:

$$\begin{aligned} MMD(\mathcal{H}, p, q) &= \sup_{f \in \mathcal{H}} \left(E_p(f(x)) - E_q(f(y)) \right) \\ &= \sup_{f \in \mathcal{H}} \left(\langle \mu_p, f \rangle_{\mathcal{H}} - \langle \mu_q, f \rangle_{\mathcal{H}} \right) \\ &= \sup_{f \in \mathcal{H}} \left(\langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} \right) \\ &= \sup_{f \in \mathcal{H}} \left\| \mu_p - \mu_q \right\|_{\mathcal{H}} \left\| f \right\|_{\mathcal{H}} \\ &\propto \sup_{\|f\|_{\mathcal{H}} \leq 1} \left\| \mu_p - \mu_q \right\|_{\mathcal{H}} \\ &= \left\| \mu_p - \mu_q \right\|_{\mathcal{H}} . \end{aligned}$$

  ...hence the name maximum mean discrepancy.
- Note this allows us to equivalently think about function space (as in the previous figures) or feature space (as in μ_p, etc.).
- Choosing p as the joint and q as the product of the marginals recovers HSIC:

$$HSIC\left(p_{xy}, \mathcal{H}_x, \mathcal{H}_y\right) = \left\| \mu_{p_{xy}} - \mu_{p_x p_y} \right\|_{HS}.$$

Conditions such that MMD is a metric

- For MMD to be useful, we want $MMD(\mathcal{H}, p, q) = 0 \Leftrightarrow p = q$.

  Simple counterexample: $k(x, x') = c \Leftrightarrow MMD = 0 \forall p, q$.

- Then MMD is a metric (already has ≥ 0, symmetry, triangle inequality).
- Literature has many types of kernels:
  - Characteristic: $\mu \to \int_{\mathcal{X}} k(\cdot, x) d\mu(x)$ is injective.

    (Note this is precisely preserving the above condition, since injectivity = distinctness.)
  - Universal: k is continuous, $\mathcal{X}$ compact, and $\mathcal{H}$ dense in the space of bounded continuous functions on $\mathcal{X}$ (wrt ℓ_∞).
  - Strictly pd (the usual)
  - Conditionally strictly pd (the usual, but $v^\top 1 = 0$ in $v^\top K v > 0$.
  - Integrally strictly pd...
- Universal and characteristic kernels both have $MMD(\mathcal{H}, p, q) = 0 \Leftrightarrow p = q$.
- For radial kernels on $\mathbb{R}^d$, all above coincide.
- In particular, a stationary k with power spectral density that is nonzero everywhere (support $= \mathbb{R}^d$) $\Leftrightarrow$ k is characteristic.
- The similarities/differences are clarified in [SFL11].

Computing MMD

- Assume μ_p, μ_q exist as before. Then:

$$\begin{aligned}
MMD^2(\mathcal{H}, p, q) &= \left(\sup_{\|f\|_{\mathcal{H}} \leq 1} \left(E_p(f(x)) - E_q(f(y)) \right) \right)^2 \\
&= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\
&= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} - 2 \langle \mu_p, \mu_q \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} \\
&= \left\langle E_p(\phi(x)), E_p(\phi(x')) \right\rangle_{\mathcal{H}} \cdots \\
&= E_{xx'} \left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{H}} \cdots \\
&= E_{xx'}(k(x, x')) - 2E_{xy}(k(x, y)) + E_{yy'}(k(y, y')).
\end{aligned}$$

- Which suggests the following finite (unbiased) sample statistic:

$$MMD_u^2(\mathcal{H}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} k(x_i, x_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(y_i, y_j).$$

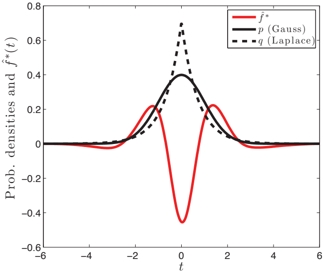
- There is a similar biased, minimum variance estimator (see [GBR$^+$12]).

Another example

$$MMD(\mathcal{H}, p, q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(E_p(f(x)) - E_q(f(y))\right) = \|\mu_p - \mu_q\|_{\mathcal{H}}.$$

- Because we used C-S, we know f is a scaled version of $\mu_p - \mu_q \in \mathcal{H}$. Thus:

$$f(x') \quad = \quad \langle \phi(x'), \alpha(\mu_p - \mu_q) \rangle_{\mathcal{H}} \quad \propto \quad E_x(k(x', x)) - E_y(k(x', y)).$$



- f is the (scaled) function achieving the supremum of the MMD objective.
- It is often called the *witness function*, as it witnesses the MMD value.

Hypothesis testing

- Using the unbiased statistic:

$$MMD_u^2(\mathcal{H}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}k(x_i,x_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}k(y_i,y_j).$$

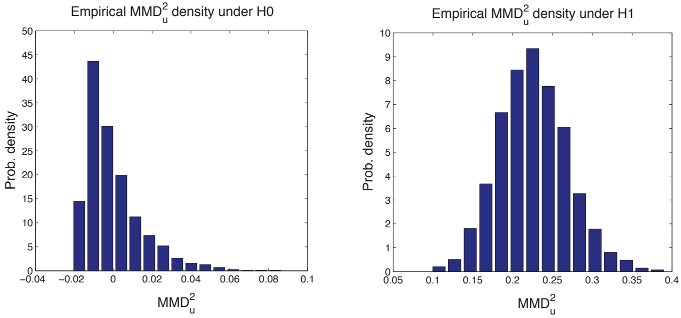
- [GBR$^+$12, Thm 10] shows that for:
  - null hypothesis $H_0 = \{p = q\}$,
  - equal sample sizes $m = n$,
  - bound $k_{max} = \max_{x,x'} k(x,x')$,
  - and power $\alpha =$ Prob(reject H_0 $|H_0$ true),
- A simple rejection test with rejection region $\mathcal{R}$ will achieve power α, where

$$\mathcal{R} = \left\{x_1, y_1, ..., x_m, y_m : MMD_u^2(\mathcal{H}) \geq \frac{1}{\sqrt{m}} 4k_{max}\sqrt{-\log\alpha}\right\}.$$

- Note these are conservative in that they don't depend on the distribution, and can be improved; see [GBR$^+$12, §5].
- MMD two-sample tests can be applied out of the box (in principle).

Another example

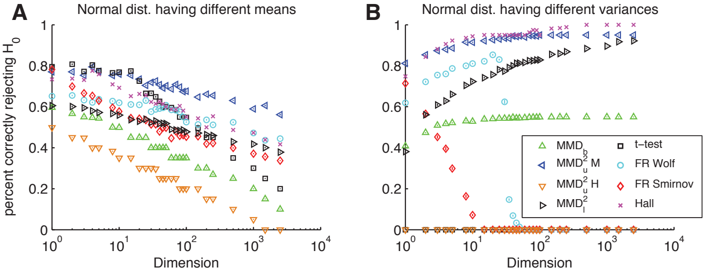
- Consider two test cases to see the distribution of MMD_u^2 under finite sampling.
- Left: test statistic under $p = q = \mathcal{N}(0,1)$; 50 samples.
- Right: test statistic under $p = Lap(0,1),\ q = Lap(0, 3\sqrt{2})$; 100 samples.



- Note: unclear (to me) which kernel (equiv, $\mathcal{H}$) is used here.

Results

- Performance separating gaussians with different means (left) and different variance (right). Test level is $\alpha = 0.05$.



- MMD_uH is the rejection region described earlier (Hoeffding bound).
- MMD_uM is an improved (but somewhat snoopy) moment-matched region.

Outline

References

[GBR+12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[GBSS05] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

[SFL11] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *The Journal of Machine Learning Research*, 12:2389–2410, 2011.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.