Kernel Hypothesis Testing and Feature Selection

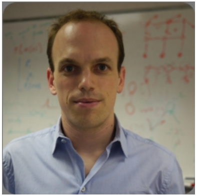
Dino Sejdinovic

Gatsby Unit, CSML, University College London

Berlin, 28 July 2014

Soumyajit De (Rahul)

Heiko Strathmann



Arthur Gretton

- **GSoC'12** (**Heiko**): Large-scale Two-Sample tests and kernel selection
- **GSoC'14** (**Rahul**): Block-based Two-Sample tests, independence tests and feature selection

Detecting pairwise dependence

X_1:

Y_1: The Dandie Dinmont Terrier is a sweet and hardy dog with lots of personality and pluck. He shows incredible loyalty to his owner, and is utterly devoted to his family. He is affectionate and loves to cuddle and be held in his owner's arms. He will follow you all over the house...

X_2:

$$\overset{?}{\Longleftrightarrow}$$

Y_2: The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps. He is a clown with a sense of humor, but he is still a true terrier: determined, keen, alert, inquisitive, and spirited....

X_3:

Y_3: Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament: inquisitive, bossy, feisty, and fearless. They are intelligent and can be a bit mischievous. Warn your flowers – many Cairns love to dig! They are not usually problem barkers, but will bark if bored or lonely...

$\cdots$[from justdogbreeds.com]

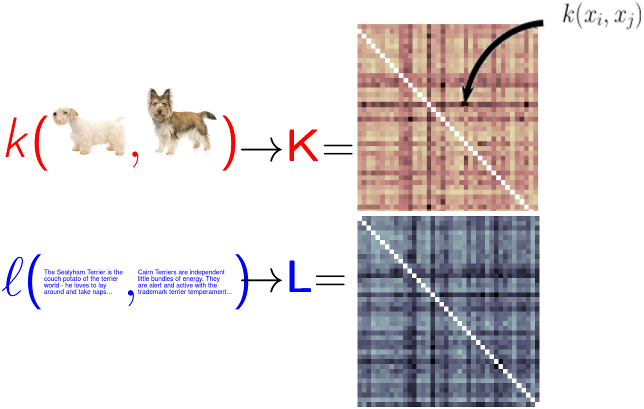
...

Detecting pairwise dependence

$k\Big($, $\Big)$

$\ell\Big($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $\Big)$

Detecting pairwise dependence

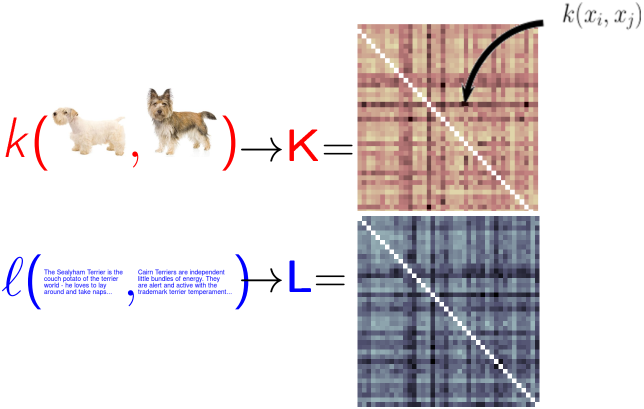


$k(x_i, x_j)$
$k(\cdot,\cdot) \rightarrow \mathbf{K} =$
$\ell(\cdot,\cdot) \rightarrow \mathbf{L} =$

The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps...

Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament...

Detecting pairwise dependence


$k(x_i, x_j)$
$k($, $) \rightarrow \mathsf{K} =$
$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $) \rightarrow \mathsf{L} =$

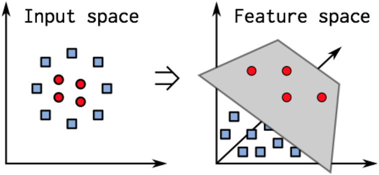
- **Idea**: measure similarity between the kernel matrices

$$\left\langle \tilde{\mathsf{K}}, \tilde{\mathsf{L}} \right\rangle = \boxed{\mathrm{Tr}\left(\tilde{\mathsf{K}}\tilde{\mathsf{L}}\right)}$$

- $\tilde{K} = HKH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (centering matrix)

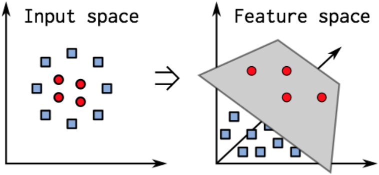
Kernel Embedding

- **feature map**: $x \mapsto k(\cdot, x) \in \mathcal{H}_k$ instead of $x \mapsto (\varphi_1(x), \ldots, \varphi_s(x)) \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$ inner products easily **computed**

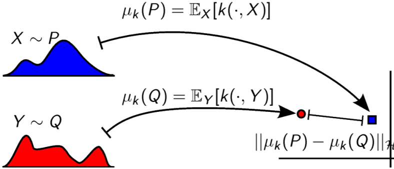


Kernel Embedding

- **feature map**: $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  instead of
  $x \mapsto (\varphi_1(x), \ldots, \varphi_s(x)) \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products easily **computed**

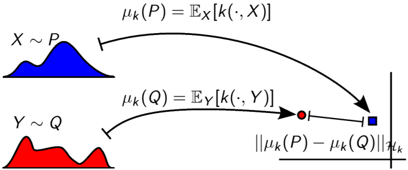


- **embedding**:
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  instead of
  $P \mapsto (\mathbb{E}\varphi_1(X), \ldots, \mathbb{E}\varphi_s(X)) \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X,Y} k(X, Y)$
  inner products easily **estimated**


$\mu_k(P) = \mathbb{E}_X[k(\cdot, X)]$
$X \sim P$
$\mu_k(Q) = \mathbb{E}_Y[k(\cdot, Y)]$
$Y \sim Q$
$\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}}$

Maximum Mean Discrepancy

- **Maximum Mean Discrepancy (MMD)** (Borgwardt et al, 2006; Gretton et al, 2007): distance between probabilities **P** and **Q**:

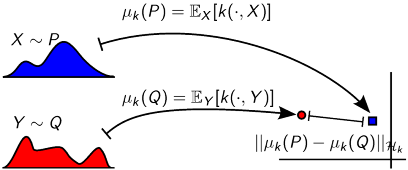


$\mu_k(P) = \mathbb{E}_X[k(\cdot, X)]$
$X \sim P$
$\mu_k(Q) = \mathbb{E}_Y[k(\cdot, Y)]$
$Y \sim Q$
$\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$

$$\mathrm{MMD}_k^2(\mathbf{P}, \mathbf{Q}) = \|\mu_k(\mathbf{P}) - \mu_k(\mathbf{Q})\|_{\mathcal{H}_k}^2 = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} [\mathbb{E}_{X \sim \mathbf{P}} f(X) - \mathbb{E}_{Y \sim \mathbf{Q}} f(Y)]$$

Maximum Mean Discrepancy

- **Maximum Mean Discrepancy (MMD)** (Borgwardt et al, 2006; Gretton et al, 2007): distance between probabilities **P** and **Q**:



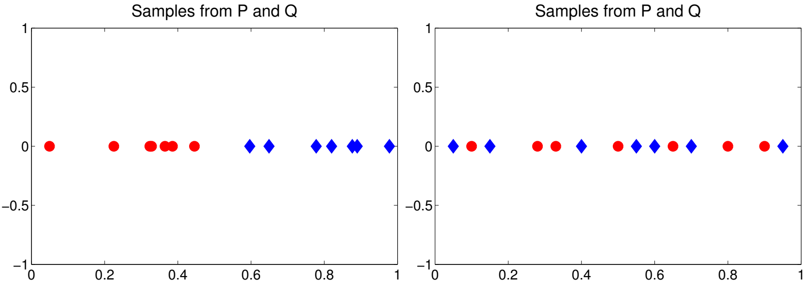
$\mu_k(P) = \mathbb{E}_X[k(\cdot, X)]$
$X \sim P$
$\mu_k(Q) = \mathbb{E}_Y[k(\cdot, Y)]$
$Y \sim Q$
$\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$

$$\mathrm{MMD}_k^2(\mathbf{P}, \mathbf{Q}) = \|\mu_k(\mathbf{P}) - \mu_k(\mathbf{Q})\|^2_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le 1} [\mathbb{E}_{X \sim \mathbf{P}} f(X) - \mathbb{E}_{Y \sim \mathbf{Q}} f(Y)]$$

- **Characteristic** kernels: $\mathrm{MMD}_k(\mathbf{P}, \mathbf{Q}) = 0$ if and only if $\mathbf{P} = \mathbf{Q}$: includes Gaussian $\exp(-\frac{1}{2\sigma^2} \|x - x'\|_2^2)$, Laplacian, Matérn etc (Sriperumbudur, 2010).

Two-Sample problem

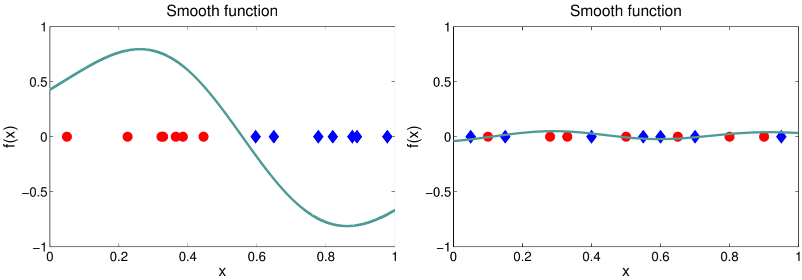
- We are given $\{x_i\}_{i=1}^{n_x} \sim \mathbf{P}$, $\{y_i\}_{i=1}^{n_y} \sim \mathbf{Q}$. Are $\mathbf{P}$ and $\mathbf{Q}$ different?



Function Showing Difference in Distributions

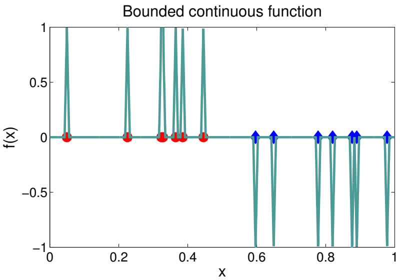
- Maximum mean discrepancy: find a **smooth function** that distinguishes **P** vs. **Q**:

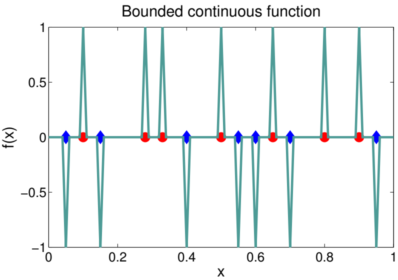
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} [\mathbb{E}_{X\sim\mathbf{P}} f(X) - \mathbb{E}_{Y\sim\mathbf{Q}} f(Y)]$$



Function Showing Difference in Distributions

- What if the “witness” is **not smooth**?





- Smoothness regulated by the choice of the kernel k, e.g., wider bandwidth in gaussian kernels implies smoother functions.

Kernel mean trick

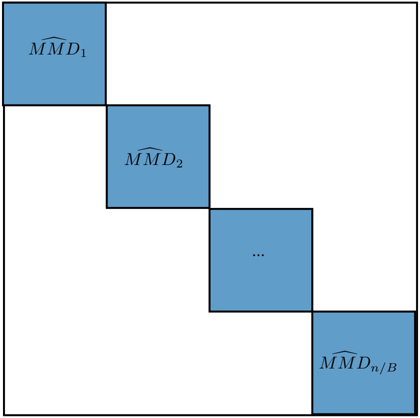
$$\mathrm{MMD}_k^2(\mathbf{P},\mathbf{Q}) = \|\mu_k(\mathbf{P}) - \mu_k(\mathbf{Q})\|_{\mathcal{H}_k}^2 = \mathbb{E}k(X,X') + \mathbb{E}k(Y,Y') - 2\mathbb{E}k(X,Y)$$

- Estimate with

$$\begin{aligned}\widehat{\mathrm{MMD}} \quad = \quad & \frac{1}{n_x(n_x-1)}\sum_{i\neq j} k(x_i,x_j) + \frac{1}{n_y(n_y-1)}\sum_{i\neq j} k(y_i,y_j) \\ & -\frac{2}{n_x n_y}\sum_{i,j} k(x_i,y_j).\end{aligned}$$

- $O\left(n^2\right)$ time to compute $\widehat{\mathrm{MMD}}$: *limited data, unlimited time*

Limited time, unlimited data



- Process blocks of size B at a time
- Complexity $O(nB)$

Statistical Hypothesis Testing

- $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$ **(null hypothesis)**
- $\mathbf{H_A}$: $\mathbf{P} \neq \mathbf{Q}$ **(alternative hypothesis)**

Statistical Hypothesis Testing

- $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$ **(null hypothesis)**
- $\mathbf{H_A}$: $\mathbf{P} \neq \mathbf{Q}$ **(alternative hypothesis)**
- Observe samples $\{x_i\}_{i=1}^{n_x} \sim \mathbf{P}$, $\{y_i\}_{i=1}^{n_y} \sim \mathbf{Q}$.

Statistical Hypothesis Testing

- $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$ **(null hypothesis)**
- $\mathbf{H_A}$: $\mathbf{P} \neq \mathbf{Q}$ **(alternative hypothesis)**
- Observe samples $\{x_i\}_{i=1}^{n_x} \sim \mathbf{P}$, $\{y_i\}_{i=1}^{n_y} \sim \mathbf{Q}$.
- Compute the value of the statistic $\widehat{\mathrm{MMD}}$ and if $\widehat{\mathrm{MMD}}$ is:
  - "*further from zero than what can be attributed to chance*": reject $\mathbf{H_0}$
  - *otherwise*: do not reject $\mathbf{H_0}$

Testing for independence via embeddings

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y$ **(null hypothesis)**
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y$ **(alternative hypothesis)**

Testing for independence via embeddings

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} = \mathbf{P}_X \mathbf{P}_Y$ **(null hypothesis)**
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} \neq \mathbf{P}_X \mathbf{P}_Y$ **(alternative hypothesis)**

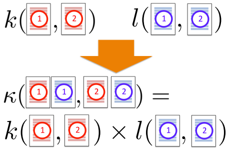
Testing for independence via embeddings

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y \Leftrightarrow \mathsf{P}_{XY} = \mathsf{P}_X \mathsf{P}_Y$ **(null hypothesis)**
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y \Leftrightarrow \mathsf{P}_{XY} \neq \mathsf{P}_X \mathsf{P}_Y$ **(alternative hypothesis)**

- **Hilbert-Schmidt Independence Criterion (HSIC)**

  Gretton et al (2005, 2008); Smola et al (2007):

  $$\|\mu_\kappa(\mathsf{P}_{XY}) - \mu_\kappa(\mathsf{P}_X \mathsf{P}_Y)\|^2_{\mathcal{H}_\kappa}$$



Testing for independence via embeddings

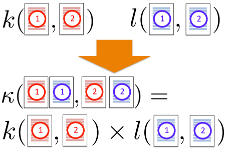
- **$\mathbf{H_0}$: $X \perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} = \mathbf{P}_X \mathbf{P}_Y$ (null hypothesis)**
- **$\mathbf{H_A}$: $X \not\!\perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} \neq \mathbf{P}_X \mathbf{P}_Y$ (alternative hypothesis)**

- **Hilbert-Schmidt Independence Criterion (HSIC)**

  Gretton et al (2005, 2008); Smola et al (2007):

  $$\|\mu_\kappa(\mathbf{P}_{XY}) - \mu_\kappa(\mathbf{P}_X \mathbf{P}_Y)\|^2_{\mathcal{H}_\kappa}$$

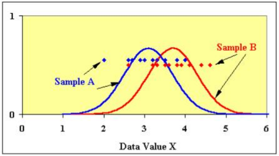
- Empirical HSIC= $\frac{1}{n^2} \boxed{\mathrm{Tr}\left(\tilde{\mathbf{K}}\tilde{\mathbf{L}}\right)}$

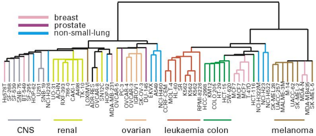


Computing the threshold

- distribution under the null hypothesis:
  $\frac{n_x n_y}{n_x+n_y}\widehat{\mathrm{MMD}} \xrightarrow{d} \sum_{r=1}^{\infty} \lambda_r \left(Z_r^2 - 1\right), \quad \{Z_r\} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
  - $\{\lambda_r\}$ depend on the kernel k and the underlying distribution $\mathbf{P}$
- Need the $(1-\alpha)$-quantile of the null distribution:
  - Fit some simple parametric form to the null distribution (no guarantees)
  - Estimate λ_r's from the data (consistent, but requires eigendecomposition of a kernel matrix)
  - **Permutation test**: merge the samples from $\mathbf{P}$ and $\mathbf{Q}$ together, split them randomly into equal proportions and recompute statistic many times, i.e., generate samples from the null

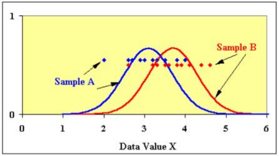
Two Sample Testing

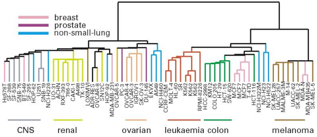




- Is there a statistically significant difference between two populations?
    - **t-tests**: Is the effect of a new drug different from placebo?

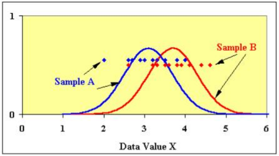
Two Sample Testing

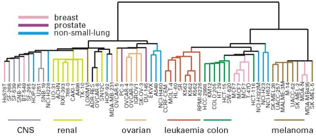




- Is there a statistically significant difference between two populations?
    - **t-tests**: Is the effect of a new drug different from placebo?
    - **Data integration**: can we train a model on data from two different sources - or should we train two separate models?

Two Sample Testing





- Is there a statistically significant difference between two populations?
  - **t-tests**: Is the effect of a new drug different from placebo?
  - **Data integration**: can we train a model on data from two different sources - or should we train two separate models?
  - **Interpreting cluster analysis**: hierarchical clustering cannot reliably distinguish between lung cancer cells and ovarian cancer cells on NCI60 dataset (Szekely & Rizzo, 2005) - is this the failure of the algorithm or is there really no difference between the two?

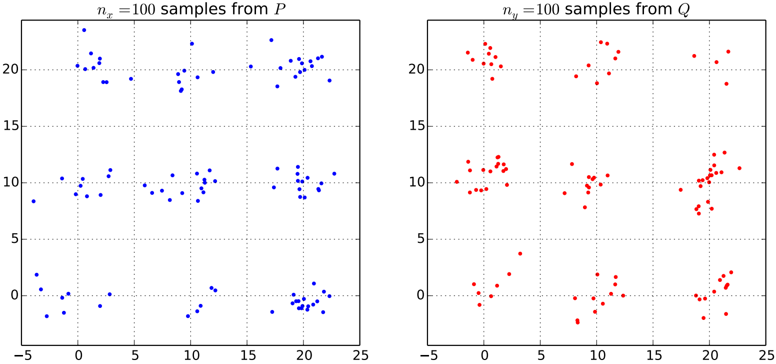
HSIC for Feature Selection

- Pick your favourite dependence measure I which is:
  - expressive enough (ideally captures nonlinear dependence)
  - easy to compute (even in high dimensions)
    - HSIC, dCor, COCO, NOCCO...
- Among the set of features $\mathcal{S} = \left\{X^{(1)}, \ldots, X^{(s)}\right\}$, pick the subset $\mathcal{T}$ of size at most $t < s$ which still contains relevant information about Y, i.e., we wish to

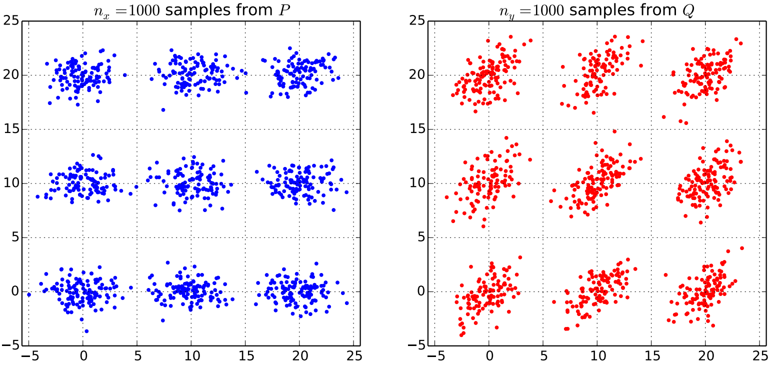
$$\text{maximize}_{\mathcal{T}\subset\mathcal{S}}\ I(\mathcal{T}; Y), \text{ subject to } |\mathcal{T}| \leq s.$$

- Forward Selection, Backward elimination...

Kernel selection: hard-to-detect differences



Kernel selection: hard-to-detect differences



- Good kernel selection crucial for the test power: scale at which the difference exists is much smaller than the overall scale of the distribution.

Other topics

- Testing for conditional independence
- Testing for multivarate interaction
- Kernel Bayes rule
- Using kernel embeddings to learn proposals in MCMC
- ???

Summary

- Kernel embeddings are awesome - computationally efficient ways to do fully nonparametric testing and inference
- Flexible and modular framework for testing and feature selection in Shogun

References

- A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf and A. Smola, **A kernel statistical test of independence**. in *Advances in Neural Information Processing Systems* 20: 585–592, MIT Press, 2008.
- A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf and A. Smola, **A Kernel Two-Sample Test**. *J. Mach. Learn. Res.* 13(Mar):723—773, 2012.
- 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* 25, Dec. 2012.
- W. Zaremba, A. Gretton and M. Blaschko, **B-test: A Non-parametric, Low Variance Kernel Two-sample Test**, in *Advances in Neural Information Processing Systems* 26, Dec. 2013.
- D. Sejdinovic, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**. *Ann. Statist.* 41(5): 2263-2291, 2013.