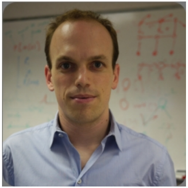
Big Hypothesis Testing with Kernel Embeddings

Dino Sejdinovic

Department of Statistics
University of Oxford

9 January 2015
UCL Workshop on the Theory of Big Data

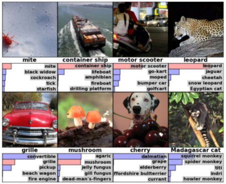
Heiko Strathmann Soumyajit De

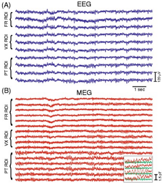


Wojciech Zaremba Matthew Blaschko Arthur Gretton

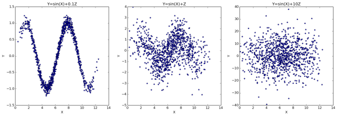
Making Hard Inference Possible

- many dimensions



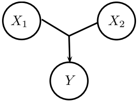


- low signal-to-noise ratio



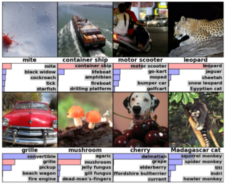
- highly non-linear assocations

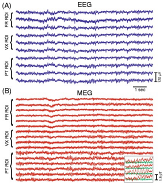
- higher-order interactions



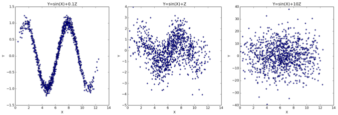
Making Hard Inference Possible

- many dimensions

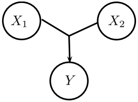




- low signal-to-noise ratio



- higher-order interactions

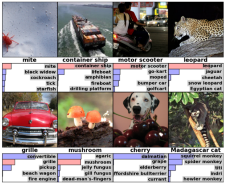


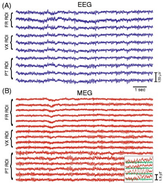
- highly non-linear assocations

need an expressive model **and** a very large number of observations

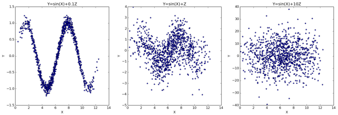
Making Hard Inference Possible

- many dimensions



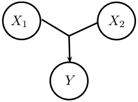


- low signal-to-noise ratio



- highly non-linear assocations

- higher-order interactions



need an expressive model **and** a very large number of observations

cannot use batch algorithms

Overview

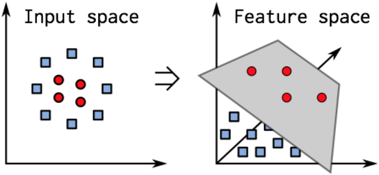
1. Kernel Embeddings and MMD
2. Scaling up Kernel Tests
3. Experiments

Outline

1. Kernel Embeddings and MMD
2. Scaling up Kernel Tests
3. Experiments

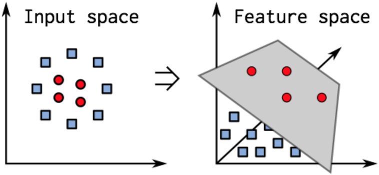
Kernel Embedding

- **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**

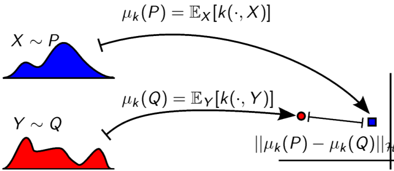


Kernel Embedding

- **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**

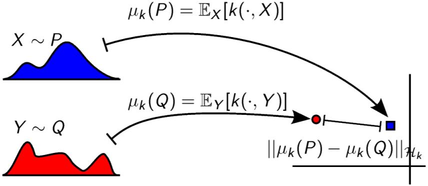


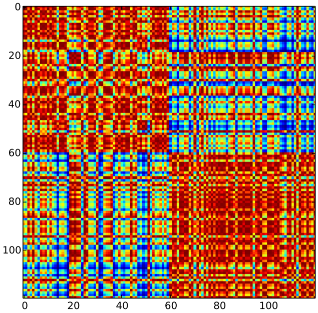
Kernel MMD (1)

Definition

Kernel metric (MMD) between P and Q:

$$\begin{aligned}\mathrm{MMD}_k^2(P,Q) &= \|\mathbb{E}k(\cdot,X)-\mathbb{E}k(\cdot,Y)\|_{\mathcal{H}_k}^2 \\ &= \mathbb{E}_{XX'}k(X,X')+\mathbb{E}_{YY'}k(Y,Y')-2\mathbb{E}_{XY}k(X,Y)\end{aligned}$$





Kernel MMD (2)

- A polynomial kernel $k(z,z') = \left(1 + z^\top z'\right)^s$ captures the difference in first s moments only
- For a certain family of kernels (**characteristic**): $\mathrm{MMD}_k(P,Q) = 0$ if and only if $P = Q$: Gaussian $\exp(-\frac{1}{2\sigma^2}\|z - z'\|_2^2)$, Laplacian, inverse multiquadratics, B_{2n+1}- splines...
- Under mild assumptions, k-MMD metrizes weak* topology on probability measures (Sriperumbudur, 2010):

$$\mathrm{MMD}_k(P_n, P) \to 0 \Leftrightarrow P_n \rightsquigarrow P$$

Nonparametric two-sample tests

- Testing $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$ vs. $\mathbf{H_A}$: $\mathbf{P} \neq \mathbf{Q}$
  based on samples $\{x_i\}_{i=1}^{n_x} \sim \mathbf{P}$, $\{y_i\}_{i=1}^{n_y} \sim \mathbf{Q}$.
- Test statistic is an estimate of
  $\mathrm{MMD}_k^2(P, Q) = \mathbb{E}_{XX'} k(X, X') + \mathbb{E}_{YY'} k(Y, Y') - 2\mathbb{E}_{XY} k(X, Y)$:

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

- $O(n^2)$ to compute ($n = n_x + n_y$)
- Degenerate U-statistic: $\frac{1}{\sqrt{n}}$-convergence to MMD under $\mathbf{H_A}$, $\frac{1}{n}$-convergence to 0 under $\mathbf{H_0}$.

Gretton et al (2009, 2012)

Nonparametric independence tests

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y$
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y$

Nonparametric independence tests

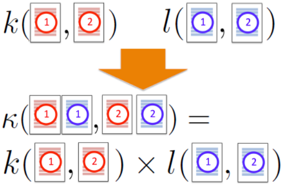
- $\mathbf{H_0}: X \perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} = \mathbf{P}_X \mathbf{P}_Y$
- $\mathbf{H_A}: X \not\perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} \neq \mathbf{P}_X \mathbf{P}_Y$

- Test statistic:

  $$\mathrm{HSIC}(X, Y) = \left\| \mu_\kappa(\hat{P}_{XY}) - \mu_\kappa(\hat{P}_X \hat{P}_Y) \right\|^2_{\mathcal{H}_\kappa},$$

  with $\kappa = k \otimes l$

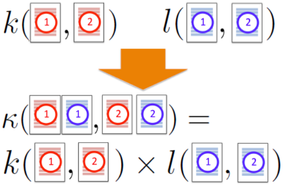
  Gretton et al (2005, 2008); Smola et al (2007)



Nonparametric independence tests

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} = \mathbf{P}_X \mathbf{P}_Y$
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}_{XY} \neq \mathbf{P}_X \mathbf{P}_Y$

- Test statistic:
  $$\mathrm{HSIC}(X, Y) = \left\| \mu_\kappa(\hat{P}_{XY}) - \mu_\kappa(\hat{P}_X \hat{P}_Y) \right\|^2_{\mathcal{H}_\kappa},$$
  with $\kappa = k \otimes l$
  Gretton et al (2005, 2008); Smola et al (2007)

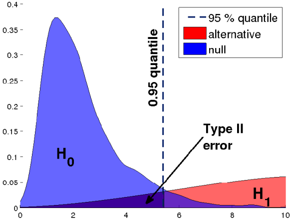


Extensions: conditional independence testing (Fukumizu, Gretton, Sun and Schölkopf, 2008; Zhang, Peters, Janzing and Schölkopf, 2011), three-variable interaction (DS, Gretton and Bergsma, 2013)

Outline

Test threshold

- under $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$:
$$\frac{n_x n_y}{n_x + n_y} \widehat{\mathrm{MMD}}_k \rightsquigarrow \sum_{r=1}^{\infty} \lambda_r \left(Z_r^2 - 1\right), \quad \{Z_r\} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$
  - $\{\lambda_r\}$ depend on both k and $\mathbf{P}$.
- expensive threshold computation:
  - Estimate leading λ_r's (requires eigendecomposition of the kernel matrix): $O(n^3)$
  - Permutation test: #shuffles $\times$ $O(n^2)$



Limited data, unlimited time

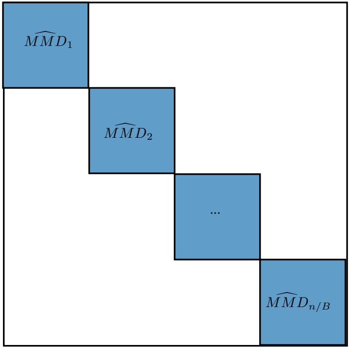
$$\mathrm{MMD}_k^2(P,Q) = \mathbb{E}_{XX'}k(X,X') + \mathbb{E}_{YY'}k(Y,Y') - 2\mathbb{E}_{XY}k(X,Y)$$

- Estimate with

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

- Complexity: $O\left(n^2\right)$.

Limited time, unlimited data



- Process mini-batches of size B at a time: $\hat{\eta}_k = \frac{B}{n} \sum_{b=1}^{n/B} \widehat{MMD}_{k,b}$
- Complexity: $O(nB)$.
- Provided $B/n \to 0$: $\frac{1}{\sqrt{n}}$-convergence to MMD if MMD $\neq 0$, $\frac{1}{\sqrt{nB}}$-convergence to 0 under $\mathbf{H_0}$.

- A.Gretton, B.Sriperumbudur, DS, H.Strathmann, S.Balakrishnan, M.Pontil and K.Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, *NIPS* 2012.
- W.Zaremba, A.Gretton, M.Blaschko, **B-test: A Non-Parametric, Low Variance Kernel Two-Sample Test**, *NIPS* 2013.

Full statistic vs. mini-batch statistic

	U-statistic	mini-batch
time	$O(n^2)$	$O(nB)$
storage	$O(n^2)$	$O(B^2)$
null distribution	infinite sum of chi-squares	**normal**
computing p-value	$O(n^3)$ or #shuffles $\times$ $O(n^2)$	$O(nB)$
convergence rate	$1/n$	$1/\sqrt{nB}$

- $\frac{n_x n_y}{(n_x+n_y)^{3/2}}\sqrt{B}\hat{\eta}_k \rightsquigarrow \mathcal{N}\left(0,\sigma_k^2\right)$ under $\mathbf{H_0}$
- σ_k^2 (depends on k and **P**) can be unbiasedly estimated on each block in $O(B^2)$ time

Asymptotic efficiency criterion

- A. Gretton, B. Sriperumbudur, DS, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, *NIPS* 2012.

Proposition

For given P and Q. Let $\eta_k = MMD_k^2(P, Q)$, *and let* σ_k^2 *be the asymptotic variance of the linear-time statistic* $\hat{\eta}_k$. *Then*

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

minimizes the asymptotic Type II error probability on $\mathcal{K}$.

Asymptotic efficiency criterion

- A. Gretton, B. Sriperumbudur, DS, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, *NIPS* 2012.

Proposition

For given P and Q. Let $\eta_k = MMD_k^2(P, Q)$*, and let* σ_k^2 *be the asymptotic variance of the linear-time statistic* $\hat{\eta}_k$*. Then*

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

minimizes the asymptotic Type II error probability on $\mathcal{K}$*.*

- We only have estimates of η_k and σ_k !
- Will the kernel optimization using plug-in esimates be consistent?
- Over what families of kernels can we perform such optimization *efficiently*?

Asymptotic efficiency criterion

- A. Gretton, B. Sriperumbudur, DS, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, *NIPS* 2012.

Proposition

For given P and Q. Let $\eta_k = MMD_k^2(P, Q)$*, and let* σ_k^2 *be the asymptotic variance of the linear-time statistic* $\hat{\eta}_k$*. Then*

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

minimizes the asymptotic Type II error probability on $\mathcal{K}$*.*

- We only have estimates of η_k and σ_k !
- Will the kernel optimization using plug-in esimates be consistent? yes!
- Over what families of kernels can we perform such optimization *efficiently*? linear combinations (MKL)

Outline

Hard-to-detect differences: Gaussian blobs

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

Hard-to-detect differences: Gaussian blobs

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.

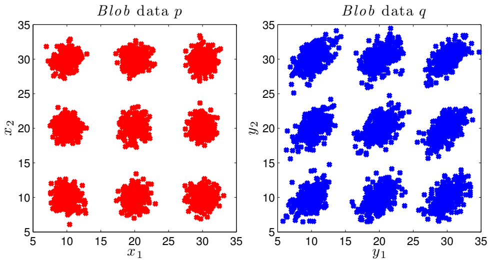


Figure : 3×3 blobs, ratio $\varepsilon = 3.2$ of largest-to-smallest eigenvalues of blobs in Q.

Gaussian blobs (2)

12×12 blobs with $\varepsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel.

Gaussian blobs (2)

12×12 blobs with $\varepsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel.

	m **per trial**	**Type II error**	**Trials**
Quadratic	5,000	$[0.7996, 0.8516]$	820
	10,000	$[0.5161, 0.6175]$	367
	$> 10,000$	Buy more RAM!	

Gaussian blobs (2)

12×12 blobs with $\varepsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel.

	m per trial	Type II error	Trials
Quadratic	5,000	$[0.7996, 0.8516]$	820
	10,000	$[0.5161, 0.6175]$	367
	$> 10,000$	Buy more RAM!	
Linear	$\sim 100,000,000$	$[0.2250, 0.3049]$	468
	$\sim 200,000,000$	$[0.1873, 0.2829]$	302
	$\vdots$	$\vdots$	$\vdots$
	$\sim 500,000,000$	$\mathbf{0.0270} \pm \mathbf{0.0302}$	111

Gaussian blobs (3)

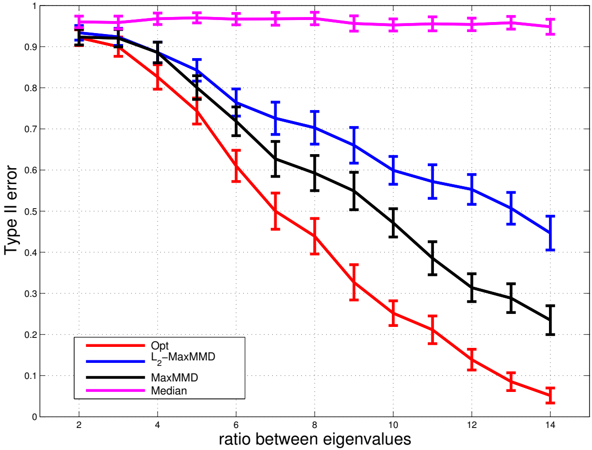


Figure : $m = 10,000$; family generated by gaussian kernels with bandwiths $\{2^{-5}, \ldots, 2^{15}\}$.

Hard-to-detect differences: UCI HIGGS

- P. Baldi, P. Sadowski, and D. Whiteson. **Searching for Exotic Particles in High-energy Physics with Deep Learning**. *Nature Communications* 5, 2014.

- benchmark dataset for distinguishing a signature of Higgs boson vs. background
- joint distributions of the azimuthal angular momenta φ for four particle jets: low-signal, low-level features
- Do *joint* angular momenta carry any discriminating information?

Hard-to-detect differences: UCI HIGGS

- P. Baldi, P. Sadowski, and D. Whiteson. **Searching for Exotic Particles in High-energy Physics with Deep Learning**. *Nature Communications* 5, 2014.

- benchmark dataset for distinguishing a signature of Higgs boson vs. background
- joint distributions of the azimuthal angular momenta φ for four particle jets: low-signal, low-level features
- Do *joint* angular momenta carry any discriminating information?

sample size:	1e4	5e4	1e5	5e5	1e6
p-value (gauss-med):	.757	.217	.475	.391	.074

Hard-to-detect differences: UCI HIGGS

- P. Baldi, P. Sadowski, and D. Whiteson. **Searching for Exotic Particles in High-energy Physics with Deep Learning**. *Nature Communications* 5, 2014.

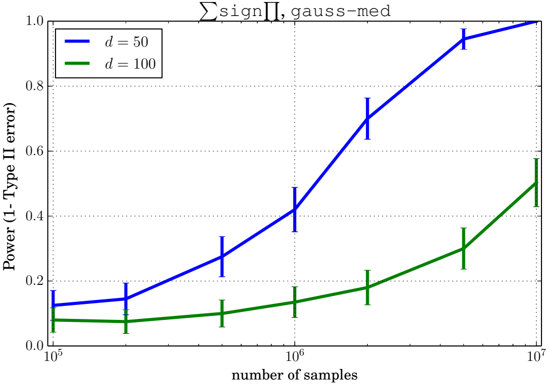
- benchmark dataset for distinguishing a signature of Higgs boson vs. background
- joint distributions of the azimuthal angular momenta φ for four particle jets: low-signal, low-level features
- Do *joint* angular momenta carry any discriminating information?

sample size:	1e4	5e4	1e5	5e5	1e6
p-value (gauss-med):	.757	.217	.475	.391	.074

train/test size:	2e3/8e3	1e4/4e4	2e4/8e4	1e5/4e5	2e5/8e5
p-value (gauss-opt):	.139	.476	.035	**6.12e-5**	**1.02e-18**

Experiment: Independence Test ($\sum \texttt{sign} \prod$)

- $X \sim \mathcal{N}(0, I_d)$,
  $Y = \sqrt{\frac{2}{d}} \sum_{j=1}^{d/2} \text{sign}\left(X_{2j-1}X_{2j}\right)|Z_j| + Z_{\frac{d}{2}+1}$, where $Z \sim \mathcal{N}\left(0, I_{\frac{d}{2}+1}\right)$



Experiment: Independence Test (`sine` $\sum$)

- $X_1, X_2 \overset{i.i.d.}{\sim} \mathrm{Unif}[0, 2\pi]$,
  $Y = \sin(X_1 + X_2) + 10Z$, with $Z \sim \mathcal{N}(0, 1)$.

`sine` $\sum$, $B = 100$	brown: $q = 1$	brown: opt	gauss: med	gauss: opt
$N = 5e5$, 1-Type II	$.277 \pm .059$	$\mathbf{.675 \pm .065}$	$.190 \pm .054$	$\mathbf{.740 \pm .061}$
Type I	$.035 \pm .025$	$.025 \pm .022$	$.085 \pm .039$	$.040 \pm .027$
$N = 1e6$, 1-Type II	$.460 \pm .069$	$\mathbf{.915 \pm .039}$	$.325 \pm .065$	$\mathbf{.905 \pm .041}$
Type I	$.055 \pm .032$	$.050 \pm .030$	$.025 \pm .022$	$.060 \pm .033$

Shogun

- Written in C++ with interfaces to Python, Matlab, Java, R.
- Google Summer of Code (2012, 2014).

Summary

- Hypothesis testing based on kernel embeddings reveals hard-to-detect differences between distributions and non-linear low-signal associations.

Summary

- Hypothesis testing based on kernel embeddings reveals hard-to-detect differences between distributions and non-linear low-signal associations.
- A simple mini-batch procedure allows us to run the tests on large-scale problems and on streaming data.

Summary

- Hypothesis testing based on kernel embeddings reveals hard-to-detect differences between distributions and non-linear low-signal associations.
- A simple mini-batch procedure allows us to run the tests on large-scale problems and on streaming data.
- Can select kernel parameters on-the-fly in order to explicitly maximise test power.

Summary

- Hypothesis testing based on kernel embeddings reveals hard-to-detect differences between distributions and non-linear low-signal associations.
- A simple mini-batch procedure allows us to run the tests on large-scale problems and on streaming data.
- Can select kernel parameters on-the-fly in order to explicitly maximise test power.
- Both kernel selection and testing in $O(n)$ time and $O(1)$ storage (if $B = \text{const}$).