Hypothesis Testing with Kernel Embeddings on Big and Interdependent Data

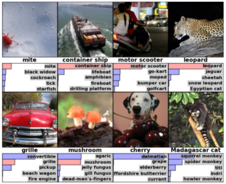
Dino Sejdinovic

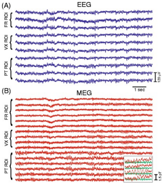
Department of Statistics
University of Oxford

6 February 2015
Department of Statistics, LSE

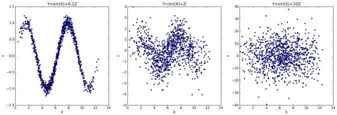
Making Hard Inference Possible

- many dimensions

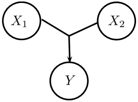




- low signal-to-noise ratio



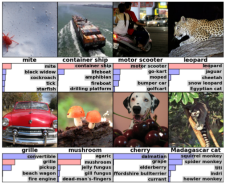
- higher-order interactions

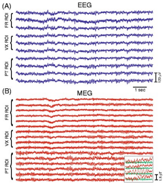


- highly non-linear assocations

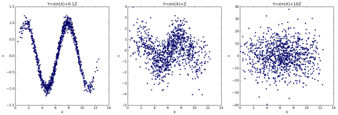
Making Hard Inference Possible

- many dimensions



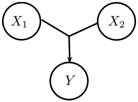


- low signal-to-noise ratio



- highly non-linear assocations

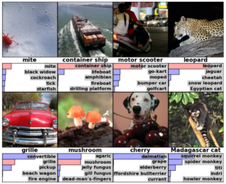
- higher-order interactions

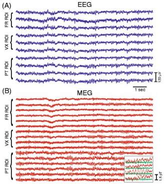


need an expressive model **and** a very large number of observations

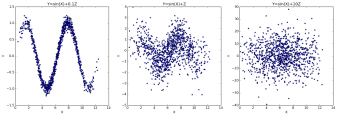
Making Hard Inference Possible

- many dimensions

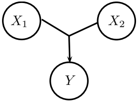




- low signal-to-noise ratio



- higher-order interactions



- highly non-linear assocations

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

cannot afford superlinear computation

Overview

1. Kernel Embeddings and MMD
2. Scaling up Kernel Tests
3. Kernel tests on time series

Outline

Reproducing Kernel Hilbert Space

RKHS

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if **evaluation functionals** $\delta_x : f \mapsto f(x)$ **are continuous** $\forall x \in \mathcal{X}$: norm convergence implies pointwise convergence.

Reproducing Kernel Hilbert Space

RKHS

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if **evaluation functionals** $\delta_x : f \mapsto f(x)$ **are continuous** $\forall x \in \mathcal{X}$: norm convergence implies pointwise convergence.

Reproducing kernel

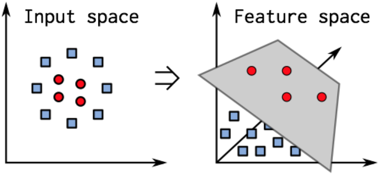
By Riesz theorem, a continuous δ_x has a representer denoted k_x s.t. $\langle f, k_x \rangle_{\mathcal{H}} = f(x)$. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by $k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$ is called a **reproducing kernel** of $\mathcal{H}$: $k_x = k(\cdot, x)$.

Moore-Aronszajn Theorem

Every positive definite function is a reproducing kernel of some $\mathcal{H}$.

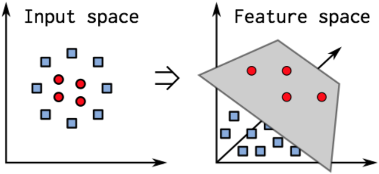
Kernel Embedding

- **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**

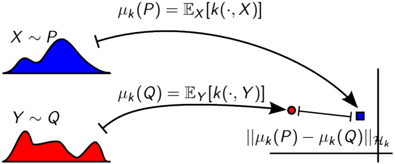


Kernel Embedding

- **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**

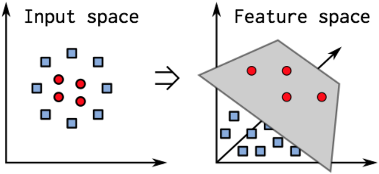


- **embedding**: $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$ instead of $P \mapsto (\mathbb{E}\varphi_1(X), \ldots, \mathbb{E}\varphi_s(X)) \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X,Y} k(X, Y)$ inner products easily **estimated**

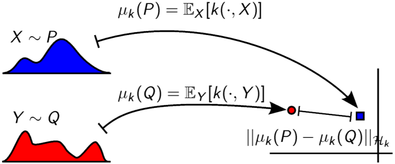


Kernel Embedding

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



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



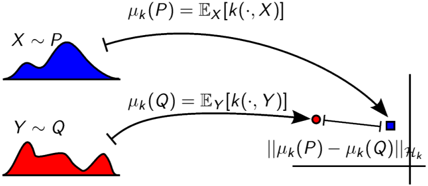
- $\mu_k(P)$ **represents expectations** w.r.t. P, i.e.,
  $\mathbb{E}_X f(X) = \mathbb{E}_X \langle f, k(\cdot, X)\rangle_{\mathcal{H}_k} = \langle f, \mu_k(P)\rangle_{\mathcal{H}_k} \;\forall f \in \mathcal{H}_k$

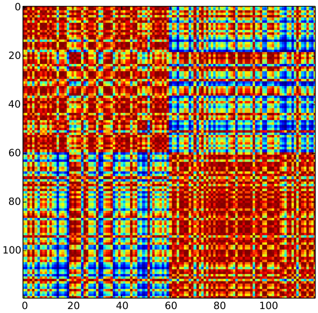
Kernel MMD (1)

Definition

Kernel metric (MMD) between P and Q:

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



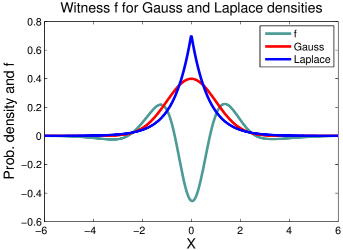


MMD as an integrable probability metric

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

$$\sup_{\substack{f\in\mathcal{H}_k,\\ \|f\|_{\mathcal{H}_k}\le 1}} \left[\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y)\right] = \sup_{\substack{f\in\mathcal{H}_k,\\ \|f\|_{\mathcal{H}_k}\le 1}} \langle f, \mu_k(P) - \mu_k(Q)\rangle_{\mathcal{H}_k}$$
$$= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}.$$

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



Kernel MMD (2)

- A polynomial kernel $k(x, x') = \left(1 + x^\top x'\right)^s$ on $\mathbb{R}^p$ captures the difference in first s (mixed) moments only
- For a certain family of kernels (**characteristic**): $\text{MMD}_k(P, Q) = 0$ iff $P = Q$: 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):

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

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

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

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

Gretton et al (2009, 2012), Lloyd & Ghahramani (2014)

Nonparametric independence tests

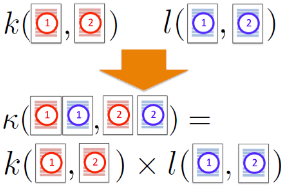
- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y$
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y$

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)

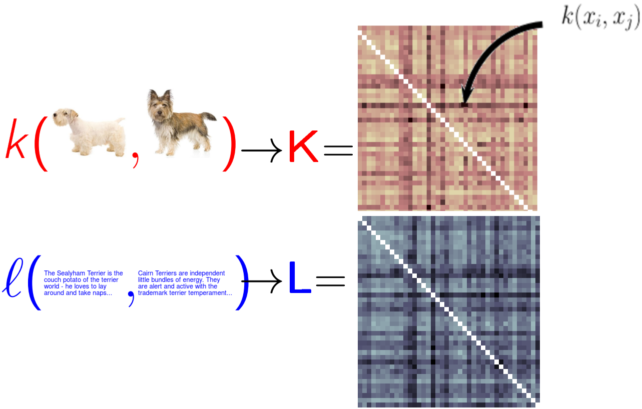


HSIC computation

$k($  $,$  $)$

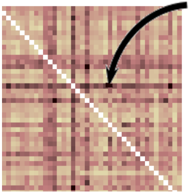
$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... $,$ Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $)$

HSIC computation

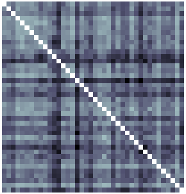


- **HSIC** measures *average similarity between the kernel matrices*: $\mathrm{HSIC}(X, Y) = \frac{1}{n^2} \langle H\mathbf{K}H, H\mathbf{L}H \rangle$
- $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (centering matrix)

HSIC computation

$k($  ,  $) \rightarrow \mathbf{K} =$ 

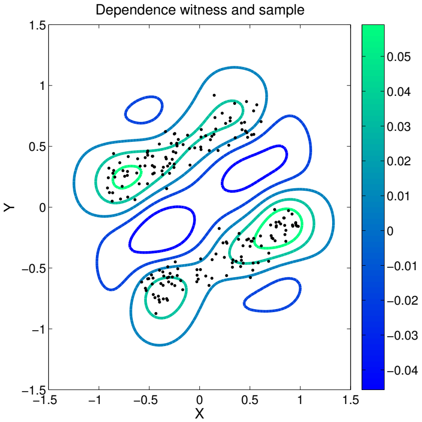
$k(x_i, x_j)$

$\ell($ 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... $) \rightarrow \mathbf{L} =$ 

- **HSIC** measures *average similarity between the kernel matrices*: $\mathrm{HSIC}(X, Y) = \frac{1}{n^2} \langle H\mathbf{K}H, H\mathbf{L}H \rangle$
- $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (centering matrix)

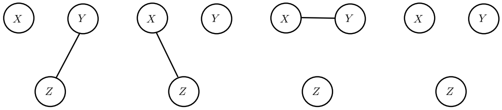
Extensions: conditional independence testing (Fukumizu, Gretton, Sun and Schölkopf, 2008; Zhang, Peters, Janzing and Schölkopf, 2011), three-variable interaction (DS, Gretton and Bergsma, 2013)

HSIC as integral probability metric

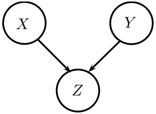


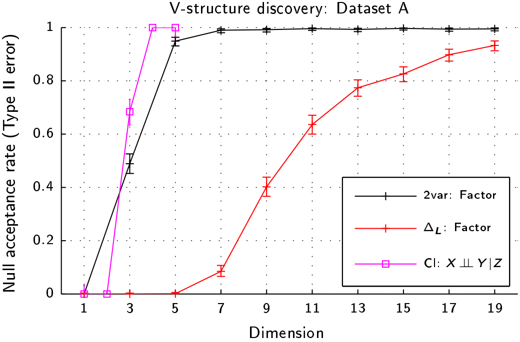
- $\|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|_{\mathcal{H}_\kappa} = \sup_f \left[\mathbb{E}_{XY} f(X,Y) - \mathbb{E}_X \mathbb{E}_Y f(X,Y)\right]$
- witness lies in the unit ball of $\mathcal{H}_\kappa = \mathcal{H}_k \otimes \mathcal{H}_l$, the RKHS of functions on $\mathcal{X} \times \mathcal{Y}$

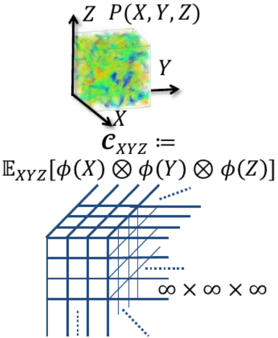
Three-variable interaction and V-structure discovery



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



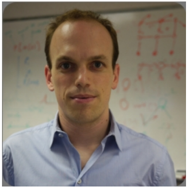




- DS, A.Gretton and W. Bergsma, **A kernel test for three-variable interactions**, ***NIPS*** **2013.**

Outline

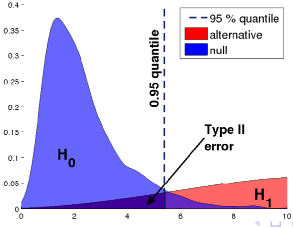
Heiko Strathmann Soumyajit De



Wojciech Zaremba Matthew Blaschko Arthur Gretton

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 $\mathbf{P}$: eigenvalues of $\mathbf{T} : L_2 \to L_2$,
  $$(\mathbf{T}f)(x) \mapsto \int f(x') \underbrace{\tilde{k}(x,x')}_{\text{centred}} d\mathbf{P}(x').$$
- expensive threshold computation:
  - Estimate leading λ_r's (eigendecomposition of the kernel matrix): $O(n^3)$
  - Permutation test: #shuffles $\times$ $O(n^2)$



Limited data, unlimited time

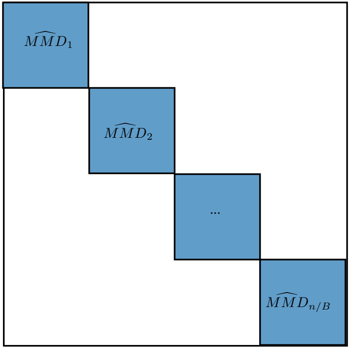
$$\text{MMD}_k^2(P,Q) = \mathbb{E}_{XX'}k(X,X') + \mathbb{E}_{YY'}k(Y,Y') - 2\mathbb{E}_{XY}k(X,Y)$$

- Estimate with

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

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

Limited time, unlimited data



- Process mini-batches of size $B = B_x + B_y$ at a time: $\hat{\eta}_k = \frac{B}{n}\sum_{b=1}^{n/B} \widehat{MMD}_{k,b}$
- Complexity: $O(nB) = \frac{n}{B} \times O(B^2)$
- Provided $B/n \to 0$: $\frac{1}{\sqrt{n}}$-convergence to MMD under $\mathbf{H_A}$, $\frac{1}{\sqrt{nB}}$-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.

Null distribution

$$\frac{n_x n_y}{(n_x+n_y)^{3/2}}\sqrt{B}\hat{\eta}_k \rightsquigarrow \mathcal{N}\left(0,\sigma_k^2\right) \text{ under } \mathbf{H_0}.$$

- σ_k^2 (depends on k and $\mathbf{P}$) can be unbiasedly estimated on each block b in $O(B^2)$ time:

$$\widehat{\left(\sigma_k^2\right)}^{(b)} = \frac{2}{B(B-3)}\left[\left(\dot{\mathbf{K}}^{(b)} \circ \dot{\mathbf{K}}^{(b)}\right)_{++} + \frac{\left(\dot{\mathbf{K}}^{(b)}_{++}\right)^2}{(B-1)(B-2)} - \frac{2}{B-2}\left(\left(\dot{\mathbf{K}}^{(b)}\right)^2\right)_{++}\right],$$

where $\dot{\mathbf{K}}^{(b)} = \mathbf{K}^{(b)} - \operatorname{diag}(\mathbf{K}^{(b)})$, and A_{++} denotes the sum of all elements of matrix A.

- Alternatively, track empirical variance of $\left\{\widehat{\mathrm{MMD}}_{k,b}\right\}_{b=1}^{n/B}$.
- No need for permutation testing.

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 $\times$ $O(n^2)$	$O(nB)$
$\mathbf{H}_0$-convergence rate	$1/n$	$1/\sqrt{nB}$

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 $\eta_k = MMD_k(P, Q)$, 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 (Hodges-Lehmann) relative efficiency 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 $\eta_k = MMD_k(P, Q)$, *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 (Hodges-Lehmann) relative efficiency 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 $\eta_k = MMD_k(P, Q)$, 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 (Hodges-Lehmann) relative efficiency 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)

Hard-to-detect differences: Gaussian blobs

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

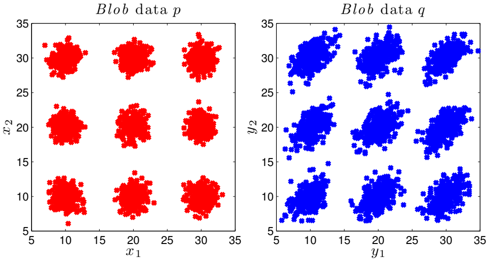


Figure : 3×3 blobs, ratio $\varepsilon = 3.2$ of largest-to-smallest eigenvalues of blobs in Q.

- Setting the bandwidth to median interpoint distance heuristic (often used in practice) "oversmooths" the distributions and misses the difference.

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
	$\vdots$	$\vdots$	$\vdots$
	$\sim 500,000,000$	$\mathbf{0.0270 \pm 0.0302}$	111

Gaussian blobs (3)

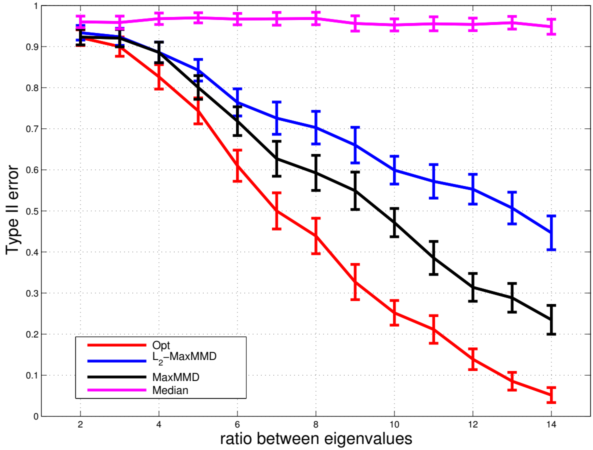


Figure : $m = 10,000$; family generated by gaussian kernels with bandwiths $\{2^{-5}, \dots, 2^{15}\}$.

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.

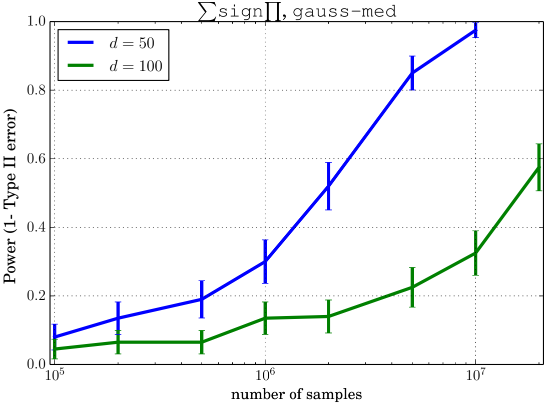
- Benchmark dataset for distinguishing a signature of Higgs boson vs. background
- Joint distributions of the azimuthal angular momenta φ for four particle jets: low-signal, low-level features
- Do *joint* angular momenta carry any discriminating information?

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 $Z \sim \mathcal{N}\left(0, I_{\frac{d}{2}+1}\right)$



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.

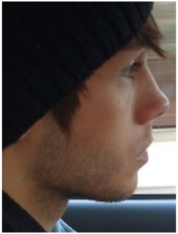
Summary

- 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 $B = \text{const}$).

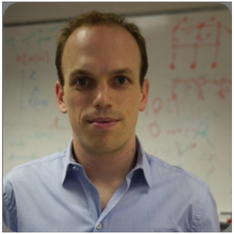
Shogun

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

Outline

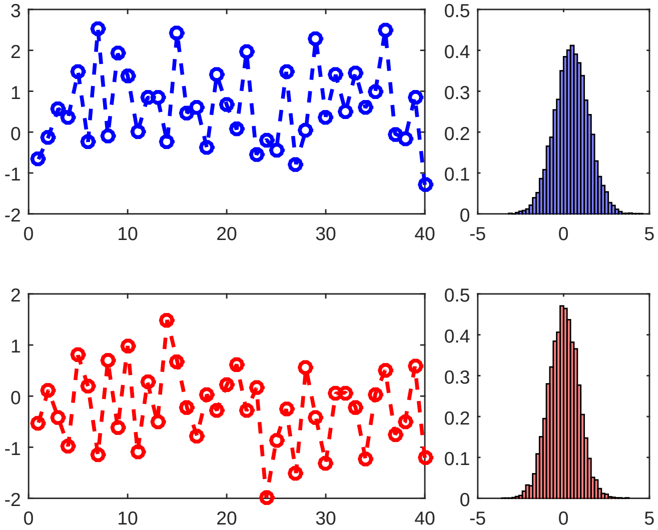


Kacper Chwialkowski



Arthur Gretton

Test calibration for dependent observations



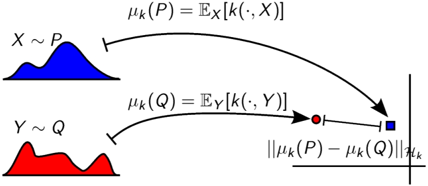
Is
P
the same
distribution as
Q
?

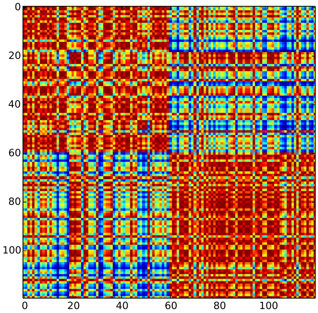
Kernel MMD

Definition

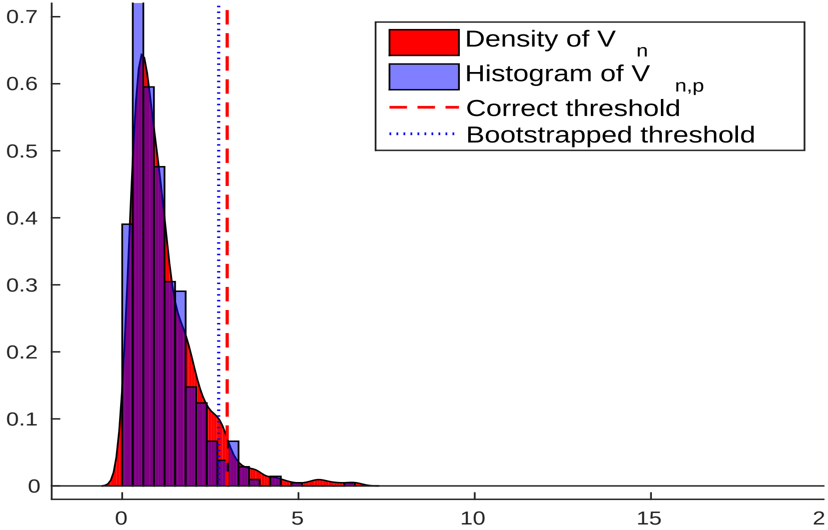
Kernel metric (MMD) between P and Q:

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



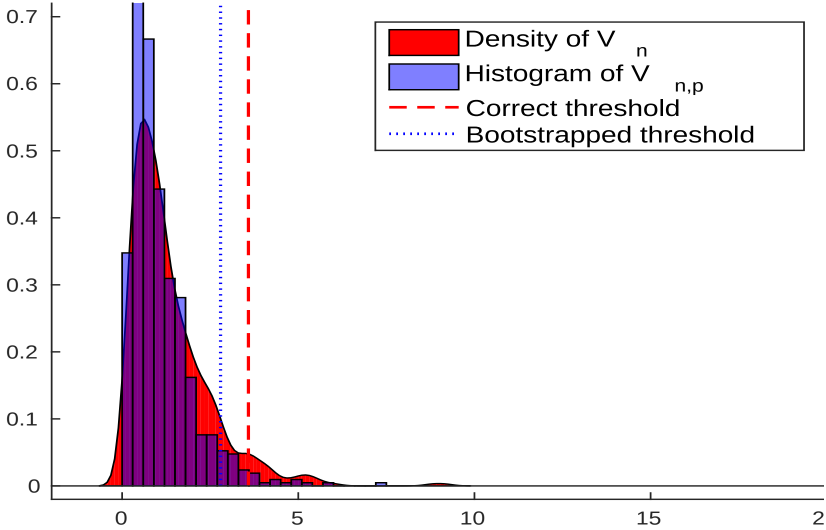


Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



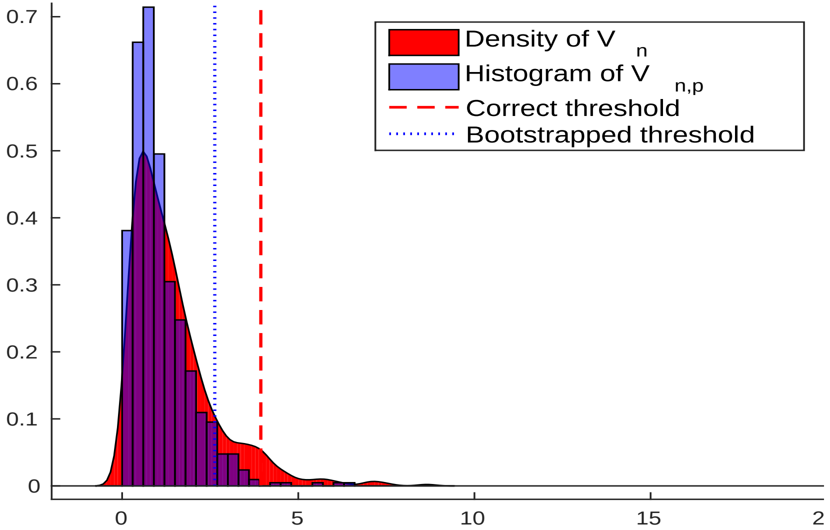
$a = 0.1$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



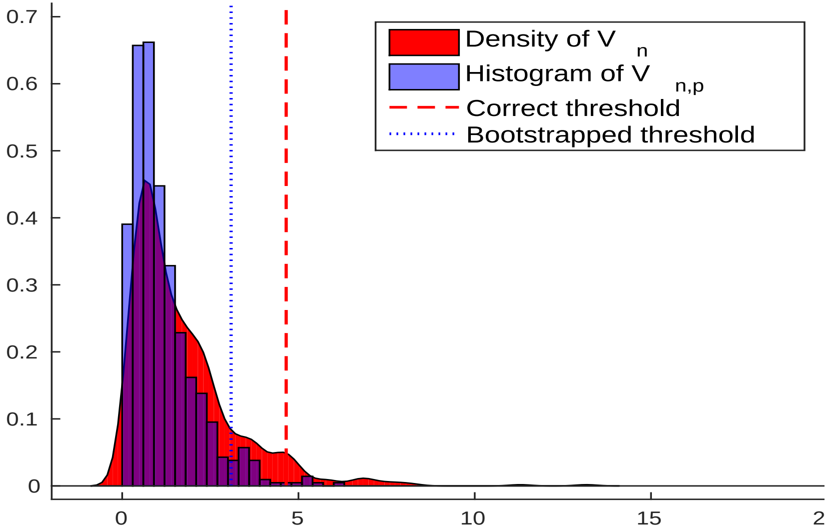
$a = 0.2$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



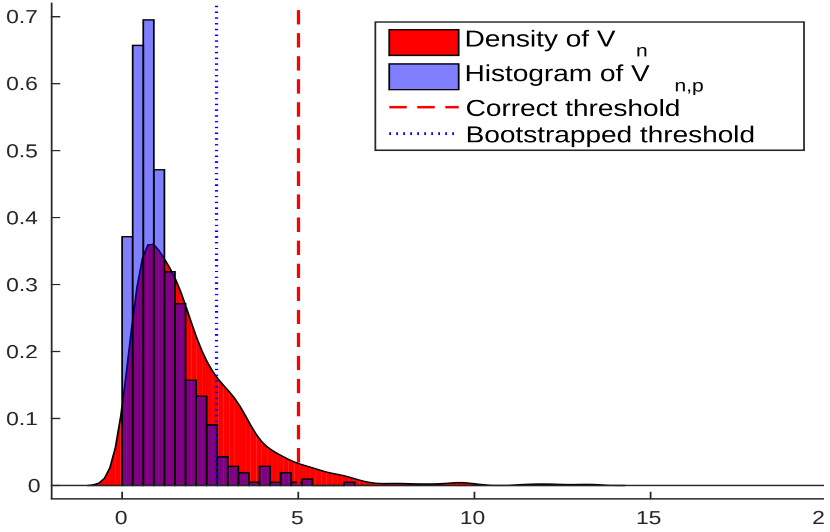
$a = 0.3$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



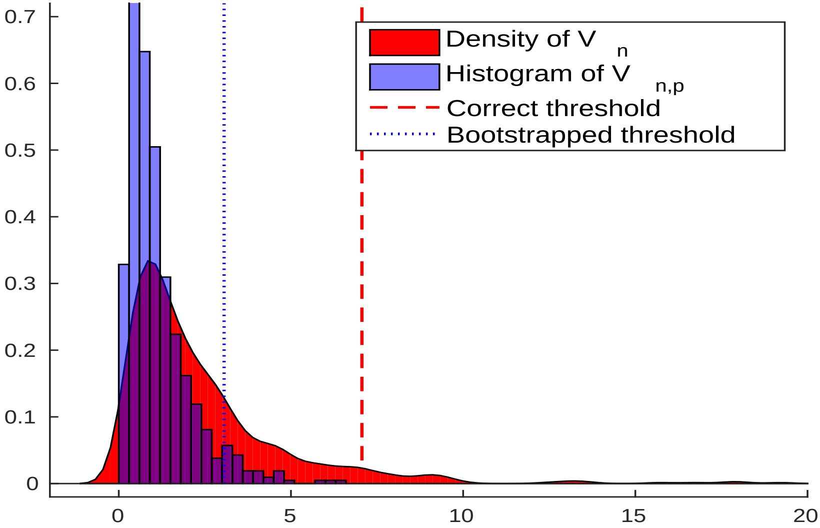
$a = 0.4$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



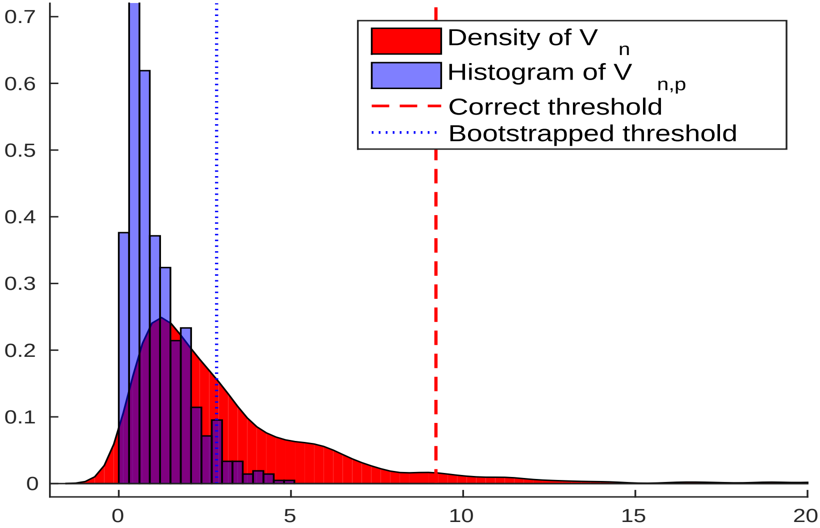
$a = 0.5$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



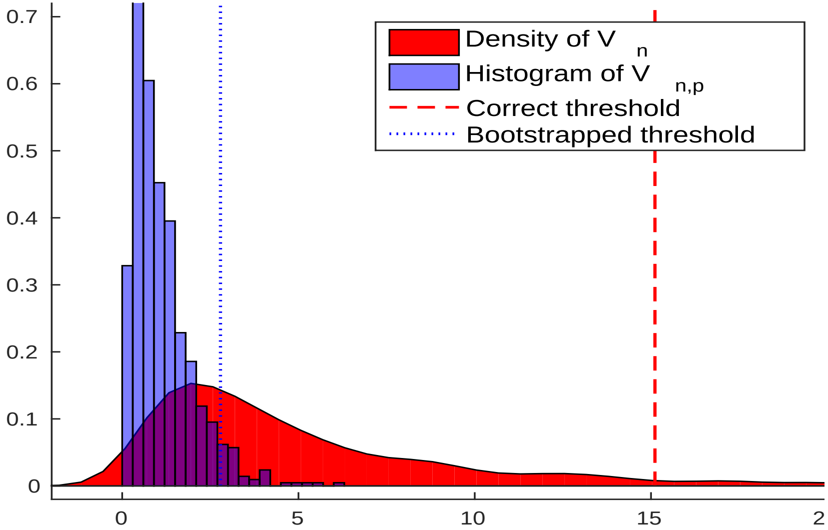
$a = 0.6$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



$a = 0.7$

Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



$a = 0.8$

Wild Bootstrap

Wild bootstrap process (Leucht and Neumann, 2013):

$W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}} \epsilon_t$ where $W_{0,n}, \epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $\tilde{W}_{t,n} = W_{t,n} - \frac{1}{n}\sum_{j=1}^{n} W_{j,n}$.

$$\widehat{\mathrm{MMD}}_{k,wb} := \frac{1}{n_x^2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_x} \tilde{W}^{(x)}_{i,n_x}\tilde{W}^{(x)}_{j,n_x} k(x_i, x_j) - \frac{1}{n_x^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y} \tilde{W}^{(y)}_{i,n_y}\tilde{W}^{(y)}_{j,n_y} k(y_i, y_j)$$
$$- \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} \tilde{W}^{(x)}_{i,n_x}\tilde{W}^{(y)}_{j,n_y} k(x_i, y_j).$$

Theorem (Chwialkowski, S. and Gretton, 2014)

Let k be bounded and Lipschitz continuous, and let $\{X_t\} \sim P$ and $\{Y_t\} \sim Q$ both be τ-dependent with $\tau(r) = O(r^{-6-\epsilon})$, but independent of each other. Then, under $\mathbf{H}_0 : P = Q$, $\varphi\left(\frac{n_x n_y}{n_x + n_y}\widehat{MMD}_k, \frac{n_x n_y}{n_x + n_y}\widehat{MMD}_{k,b}\right) \xrightarrow{P} 0$ as $n_x, n_y \to \infty$, where φ is the Prokhorov metric.

Embeddings in Mercer's Expansion

Mercer's Expansion

For a compact metric space $\mathcal{X}$, and a continous kernel k,

$$k(x, y) = \sum_{r=1}^{\infty} \lambda_r \Phi_r(x) \Phi_r(y),$$

with $\{\lambda_r, \Phi_r\}_{r \geq 1}$ eigenvalue, eigenfunction pairs of $f \mapsto \int f(x) k(\cdot, x) dP(x)$ on $L_2(P)$.

$$\begin{aligned}
\mathcal{H}_k \ni k(\cdot, x) &\leftrightarrow \left\{ \sqrt{\lambda_r} \Phi_r(x) \right\} \in \ell_2 \\
\mathcal{H}_k \ni \mu_k(P) &\leftrightarrow \left\{ \sqrt{\lambda_r} \mathbb{E} \Phi_r(X) \right\} \in \ell_2
\end{aligned}$$

$$\left\| \mu_k(\hat{P}) - \mu_k(\hat{Q}) \right\|_{\mathcal{H}_k}^2 = \sum_{r=1}^{\infty} \lambda_r \left(\frac{1}{n_x} \sum_{t=1}^{n_x} \Phi_r(X_t) - \frac{1}{n_y} \sum_{t=1}^{n_y} \Phi_r(Y_t) \right)^2$$

Wild Bootstrap

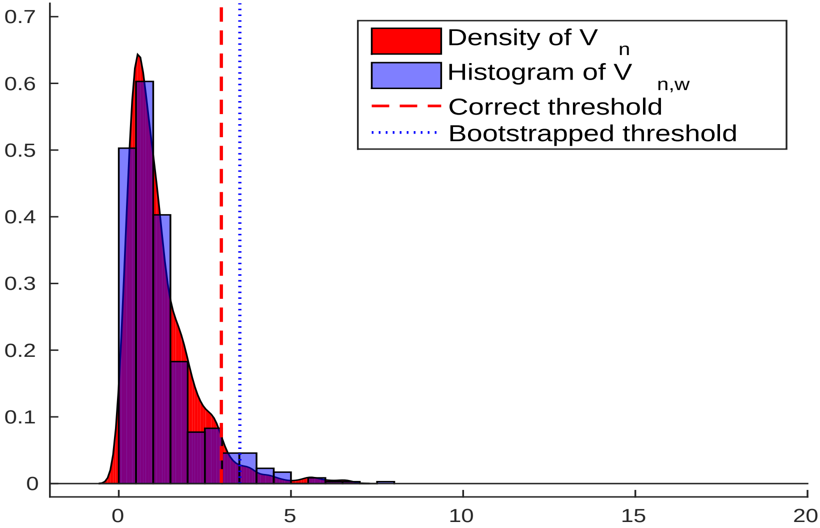
- $\rho_x = n_x/n,\ \rho_y = n_y/n$
- $\{W_{t,n}\}_{1\leq t\leq n}$, $\mathbb{E}W_{t,n} = 0$, $\mathbb{E}\left[W_{t,n}W_{t',n}\right] = \zeta\left(\frac{|t'-t|}{\ell_n}\right)$, with $\lim_{u\to 0}\zeta(u) \to 1$

$$\rho_x\rho_y n\widehat{\mathrm{MMD}}_k = \sum_{r=1}^{\infty}\lambda_r\left(\sqrt{\rho_y}\sum_{t=1}^{n_x}\frac{\Phi_r(X_t)}{\sqrt{n_x}} - \sqrt{\rho_x}\sum_{t=1}^{n_y}\frac{\Phi_r(Y_t)}{\sqrt{n_y}}\right)^2$$

$$\rho_x\rho_y n\widehat{\mathrm{MMD}}_{k,wb} = \sum_{r=1}^{\infty}\lambda_r\left(\sqrt{\rho_y}\sum_{t=1}^{n_x}\frac{\Phi_r(X_t)\tilde{W}^{(y)}_{t,n_x}}{\sqrt{n_x}} - \sqrt{\rho_x}\sum_{t=1}^{n_y}\frac{\Phi_r(Y_t)\tilde{W}^{(y)}_{t,n_y}}{\sqrt{n_y}}\right)^2$$

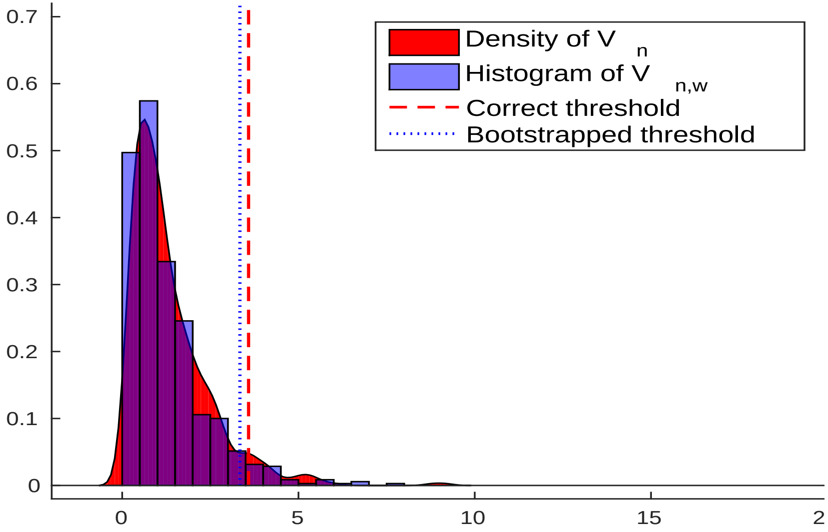
- $\mathbb{E}\left[\Phi_r(X_1)W_{1,n}\Phi_s(X_t)W_{t,n}\right] = \mathbb{E}\left[\Phi_r(X_1)\Phi_s(X_t)\right]\zeta\left(\frac{|t-1|}{\ell_n}\right) \underset{n\to\infty}{\longrightarrow} \mathbb{E}\left[\Phi_r(X_1)\Phi_s(X_t)\right]$, $\forall t, r, s$ provided dependence between X_1 and X_t "disappears fast enough" (a τ-mixing condition).

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



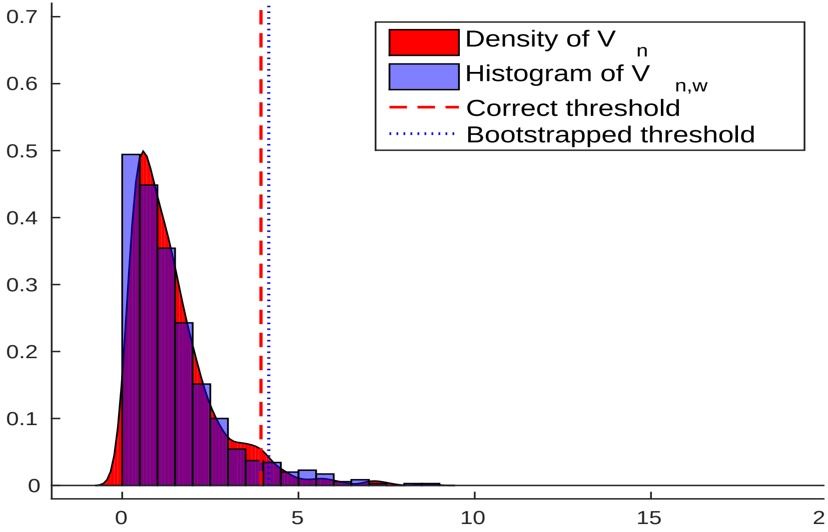
$a = 0.1$

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



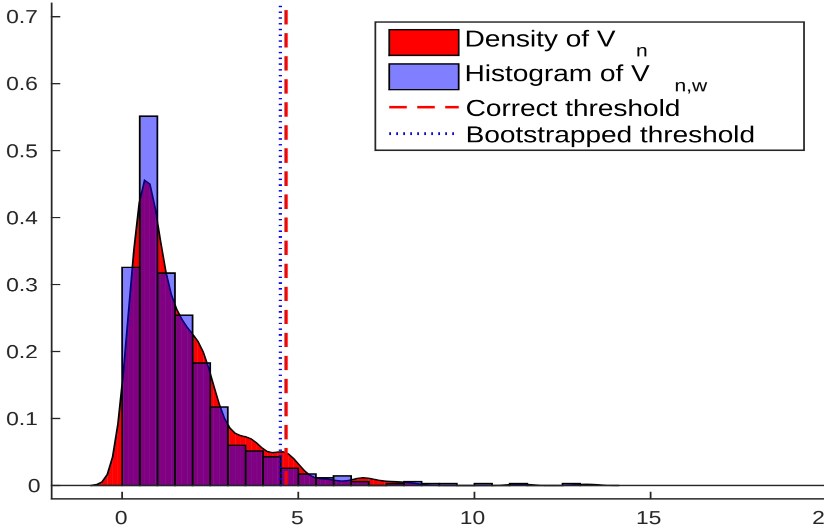
$a = 0.2$

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



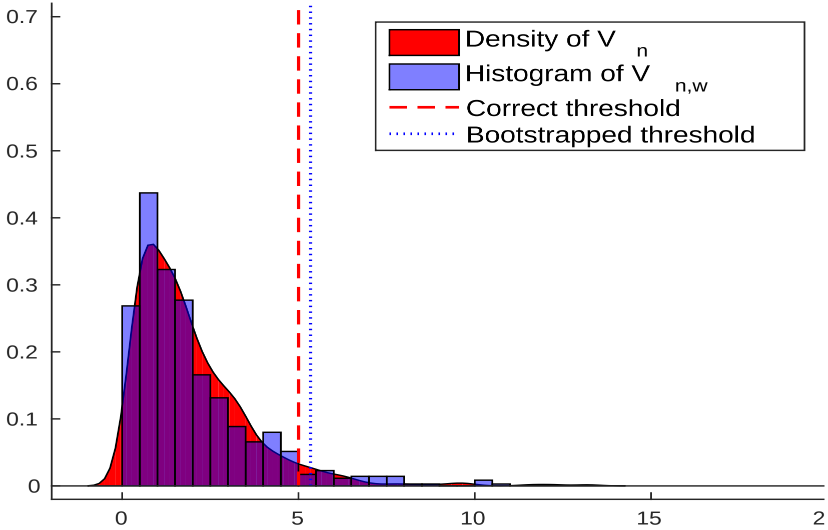
$a = 0.3$

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$



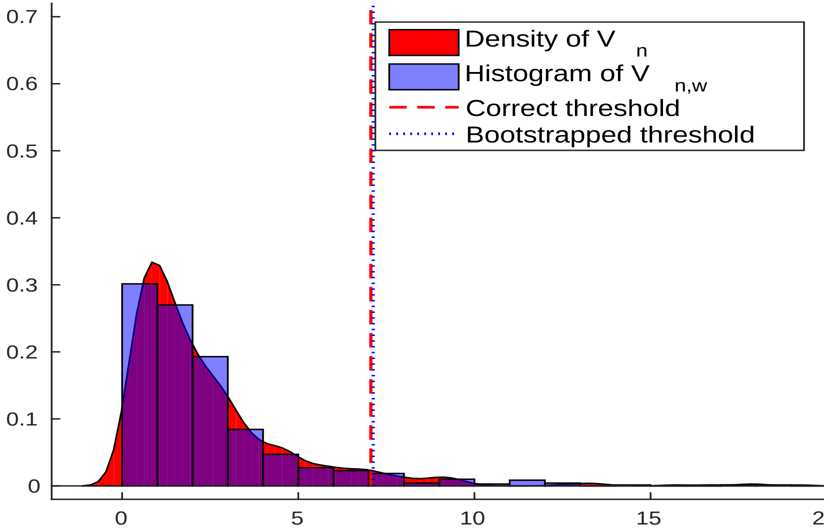
$a = 0.4$

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



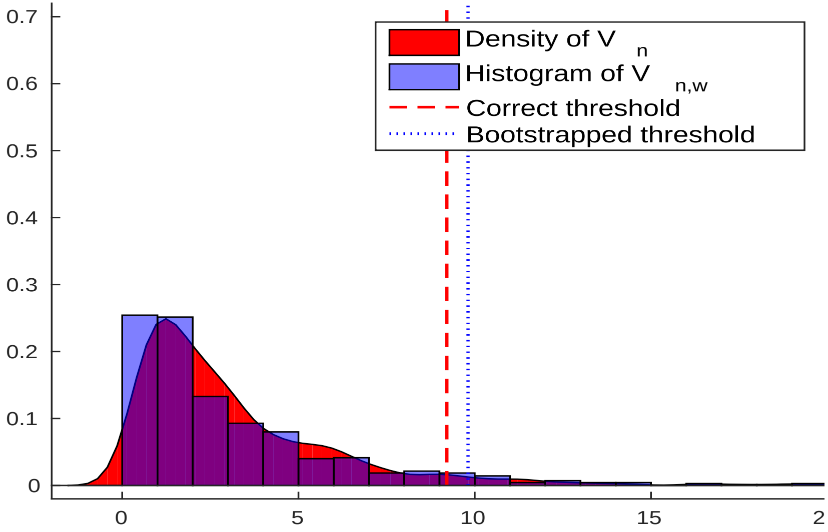
$a = 0.5$

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



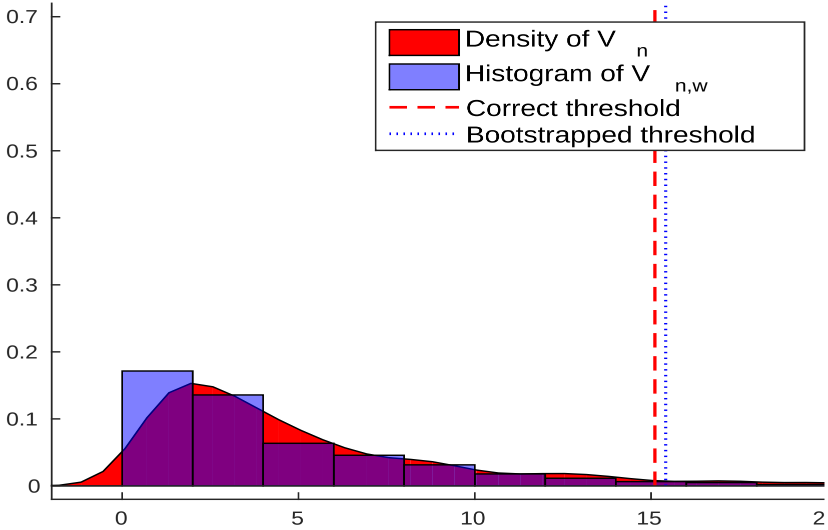
$a = 0.6$

Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$



$a = 0.7$

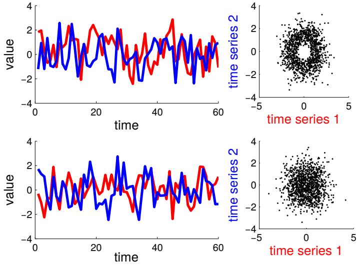
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$

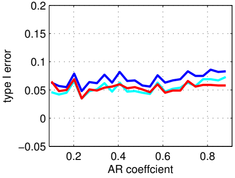


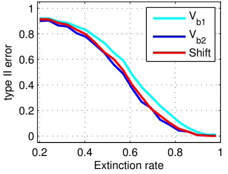
$a = 0.8$

Test calibration for dependent observations

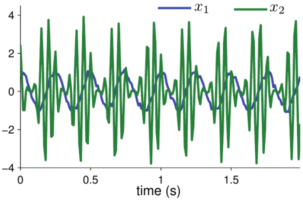
Two-sample test	experiment \ method	perm.	*wild*
MCMC diagnostics	i.i.d. vs i.i.d. ($\mathbf{H_0}$)	.040	.012
	i.i.d. vs Gibbs ($\mathbf{H_0}$)	.528	.052
	Gibbs vs Gibbs ($\mathbf{H_0}$)	.680	.060

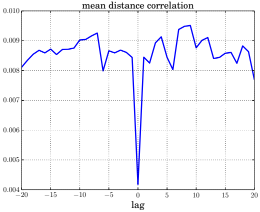


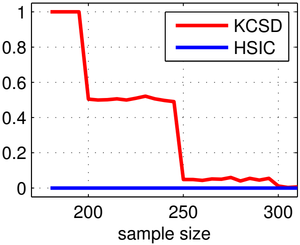




Time Series Coupled at a Lag







$$X_t = \cos(\phi_{t,1}), \qquad \phi_{t,1} = \phi_{t-1,1} + 0.1\epsilon_{1,t} + 2\pi f_1 T_s, \quad \epsilon_{1,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$

$$Y_t = [2 + C\sin(\phi_{t,1})]\cos(\phi_{t,2}), \quad \phi_{t,2} = \phi_{t-1,2} + 0.1\epsilon_{2,t} + 2\pi f_2 T_s, \quad \epsilon_{2,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1).$$

Parameters: $C = 0.4$, $f_1 = 4Hz, f_2 = 20Hz$, $\frac{1}{T_s} = 100Hz$.

- M. Besserve, N.K. Logothetis, and B. Schölkopf. **Statistical analysis of coupled time series with kernel cross-spectral density operators**. *NIPS* 2013.

Summary

- Interdependent data lead to incorrect Type I control for kernel tests (too many false positives).
- Consistency of a wild bootstrap procedure under weak long-range dependencies (τ-mixing), applicable to both two-sample and independence tests
- Applications: MCMC diagnostics, time series dependence across multiple lags

References

- K. Chwialkowski, DS and A. Gretton, **A wild bootstrap for degenerate kernel tests**. *Advances in Neural Information Processing Systems (NIPS)* 27, Dec. 2014.
- DS, A. Gretton and W. Bergsma, **A kernel test for three-variable interactions**. *Advances in Neural Information Processing Systems (NIPS)* 26, Dec. 2013.
- W.Zaremba, A.Gretton and M.Blaschko, **B-test: A Non-Parametric, Low Variance Kernel Two-Sample Test**. *Advances in Neural Information Processing Systems (NIPS)* 26, Dec. 2013.
- A. Gretton, B. Sriperumbudur, DS, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**. *Advances in Neural Information Processing Systems (NIPS)* 25, Dec. 2012.
- DS, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**. *Ann. Statist.* 41(5): 2263-2291, Oct. 2013.
- 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.

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given n, hence. . .
- . . .a much less powerful test for a given n

Linear time vs quadratic time MMD

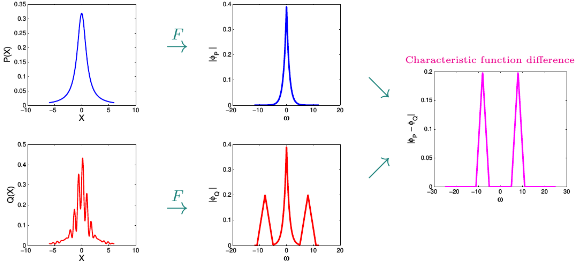
Disadvantages of linear time MMD vs quadratic time MMD

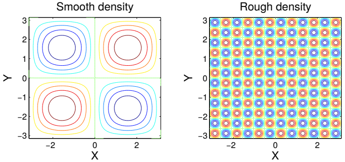
- Much higher variance for a given n, hence...
- ...a much less powerful test for a given n

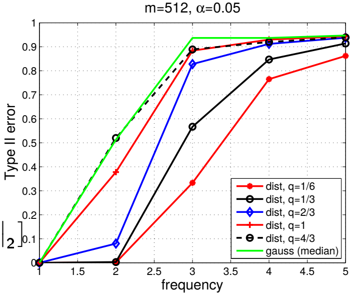
Advantages of the linear time MMD vs quadratic time MMD

- Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of χ^2)
- Both test statistic and threshold computable in $O(n)$, with storage $O(1)$ (if $B = \text{const}$).
- Given unlimited data, a given Type II error can be attained with less computation

Kernels and characteristic functions







E-distance/dCov of Székely and Rizzo (2004,2005), Székely et al (2007):

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

DS, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**. *Annals of Statistics* 41(5), p. 2263-2291, 2013.