Hypothesis Testing with Kernel Embeddings on Interdependent Data

Dino Sejdinovic

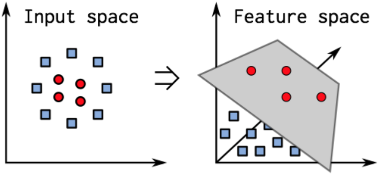
Department of Statistics
University of Oxford

joint work with
Kacper Chwialkowski and Arthur Gretton (Gatsby Unit, UCL)

9 April 2015
Dagstuhl

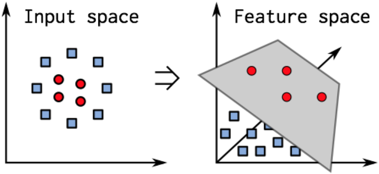
Kernel Embedding

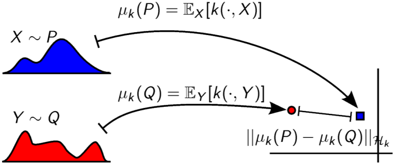
- **feature map**: $x \mapsto k(\cdot, x) \in \mathcal{H}_k$ instead of $x \mapsto (\varphi_1(x), \ldots, \varphi_s(x)) \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$ inner products easily **computed**



Kernel Embedding

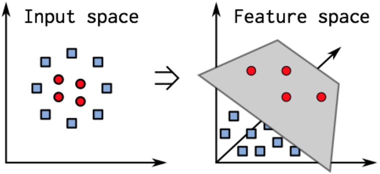
- **feature map**: $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  instead of
  $x \mapsto (\varphi_1(x), \ldots, \varphi_s(x)) \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products easily **computed**
- **embedding**:
  $P \mapsto \mu_k(P) = \mathbb{E}_{X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



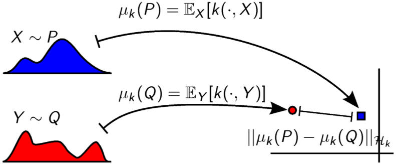


Kernel Embedding

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



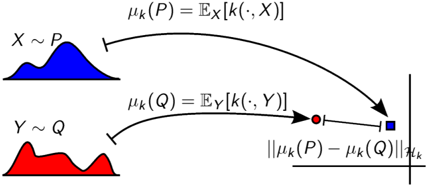
- $\mu_k(P)$ **represents expectations** w.r.t. P, i.e.,
  $\mathbb{E}_X f(X) = \mathbb{E}_X \langle f, k(\cdot, X) \rangle_{\mathcal{H}_k} = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k} \; \forall f \in \mathcal{H}_k$

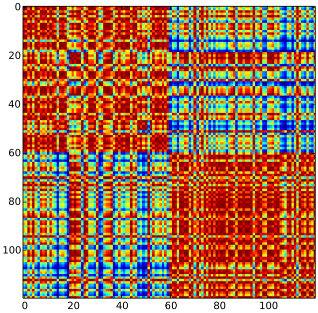
Kernel MMD

Definition

Kernel metric (MMD) between P and Q:

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





Kernel MMD

- 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/universal**): $\text{MMD}_k(P, Q) = 0$ 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
  $\mathrm{MMD}_k(P, Q) = \mathbb{E}_{XX'} k(X, X') + \mathbb{E}_{YY'} k(Y, Y') - 2\mathbb{E}_{XY} k(X, Y)$:

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

- 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 ($n = n_x + n_y$) – various approximations (block-based, random features) trade computation for power.

Gretton et al (NIPS 2009, JMLR 2012, NIPS 2012)

Test threshold

- For i.i.d. data, under $\mathbf{H_0}: \mathbf{P} = \mathbf{Q}$:
  $\frac{n_x n_y}{n_x+n_y}\widehat{\mathrm{MMD}}_k \rightsquigarrow \sum_{r=1}^{\infty} \lambda_r \left(Z_r^2 - 1\right), \quad \{Z_r\}_{r=1}^{\infty} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
  - $\{\lambda_r\}$ depend on both k and $\mathbf{P}$: eigenvalues of $\mathbf{T}: L_2 \to L_2$,

$$(\mathbf{T}f)(x) \mapsto \int f(x')\underbrace{\tilde{k}(x,x')}_{\text{centred}} d\mathbf{P}(x').$$

  - Asymptotic null distribution typically estimated using a permutation test.
- For interdependent samples, $\{Z_r\}_{r=1}^{\infty}$ are correlated, with the correlation structure dependent on the correlation structure within the samples.

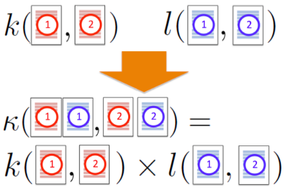
Nonparametric independence tests

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

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);
- Related to distance covariance (dCov) in statistics literature Szekely et al (AoS 2007, AoAS 2009); S. et al (AoS 2013)

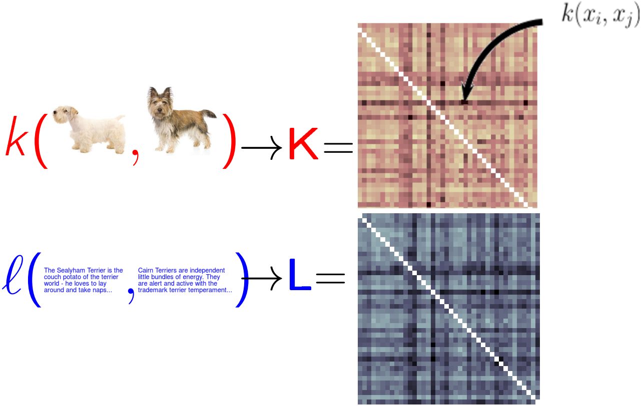


HSIC computation

$k($  ,  $)$

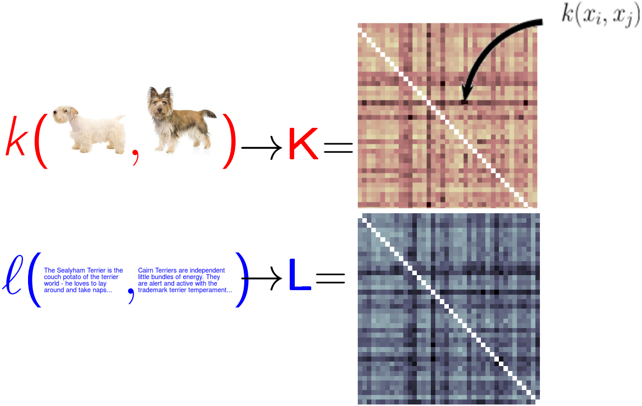
$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $)$

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)

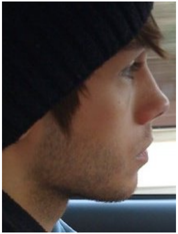
HSIC computation



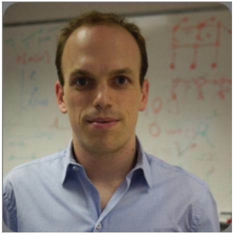
- **HSIC** measures *average similarity between the kernel matrices*: $\mathrm{HSIC}(X, Y) = \frac{1}{n^2} \langle H\mathbf{K}H, H\mathbf{L}H \rangle$
- $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (centering matrix)

Extensions: conditional independence testing (Fukumizu, Gretton, Sun and Schölkopf, 2008; Zhang, Peters, Janzing and Schölkopf, 2011), three-variable interaction / V-structure discovery (S., Gretton and Bergsma, 2013)

Kernel tests on time series

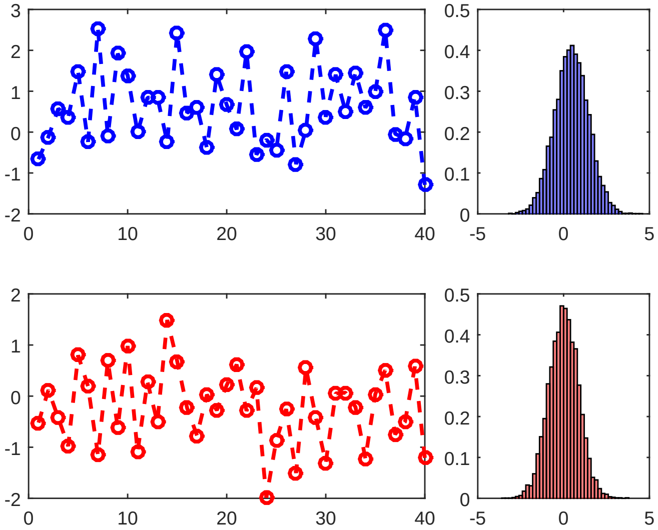


Kacper Chwialkowski



Arthur Gretton

Test calibration for dependent observations



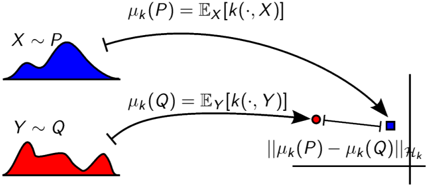
Is
P
the same
distribution as
Q
?

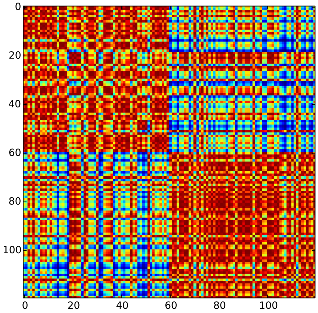
Kernel MMD

Definition

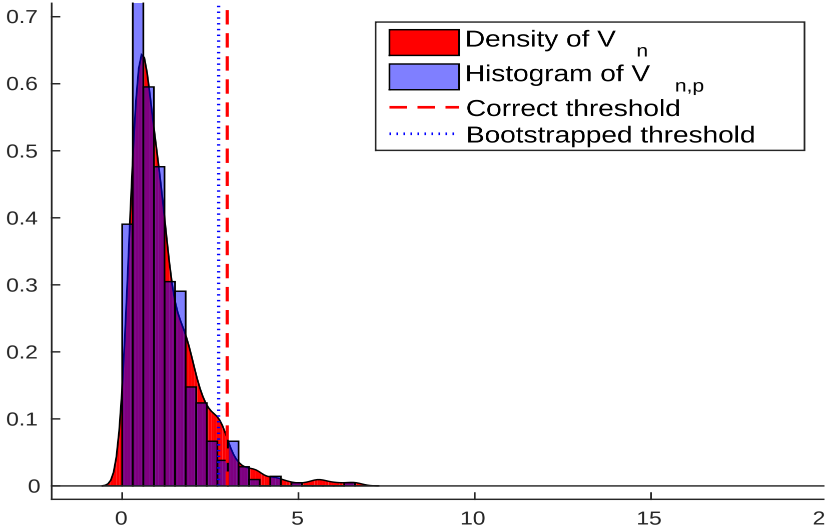
Kernel metric (MMD) between P and Q:

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



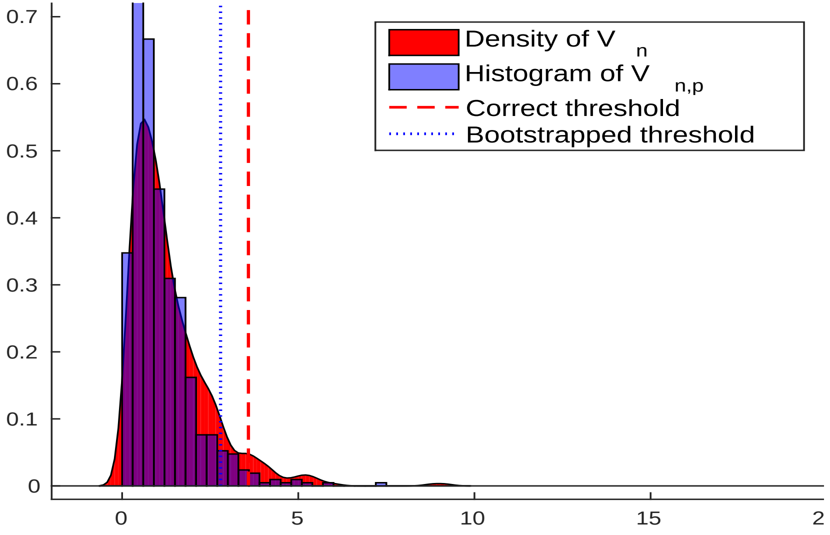


Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t$



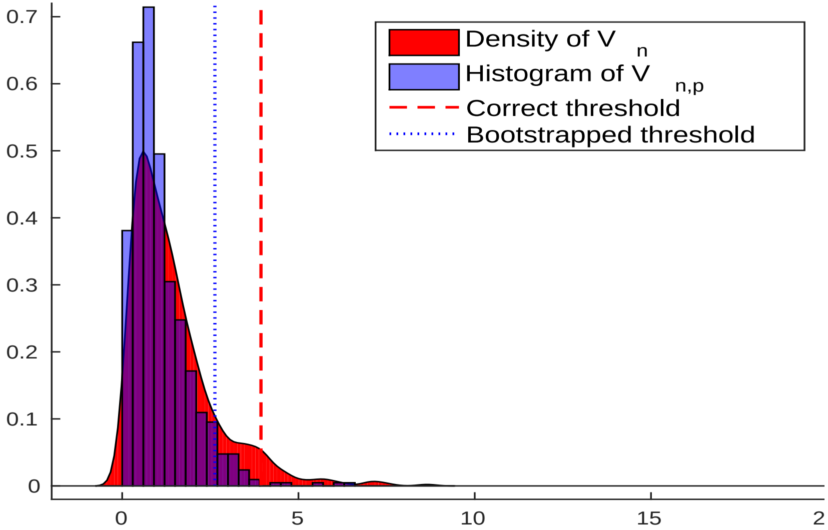
$a = 0.1$

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



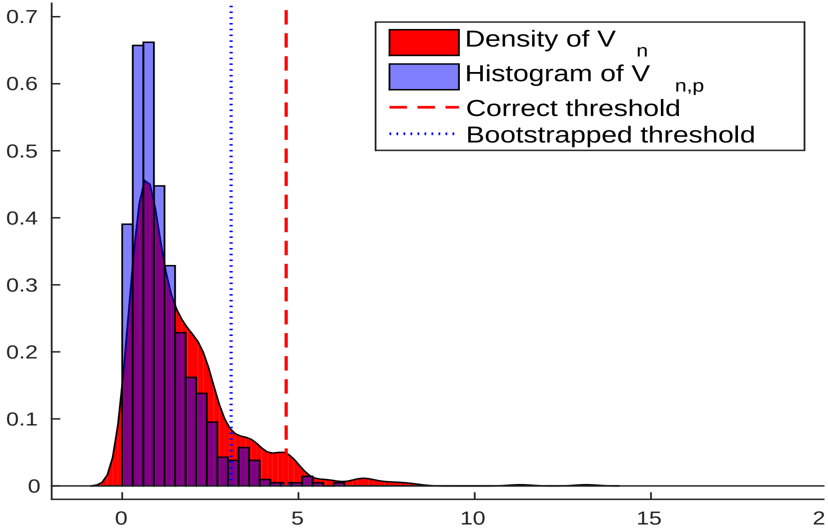
$a = 0.2$

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



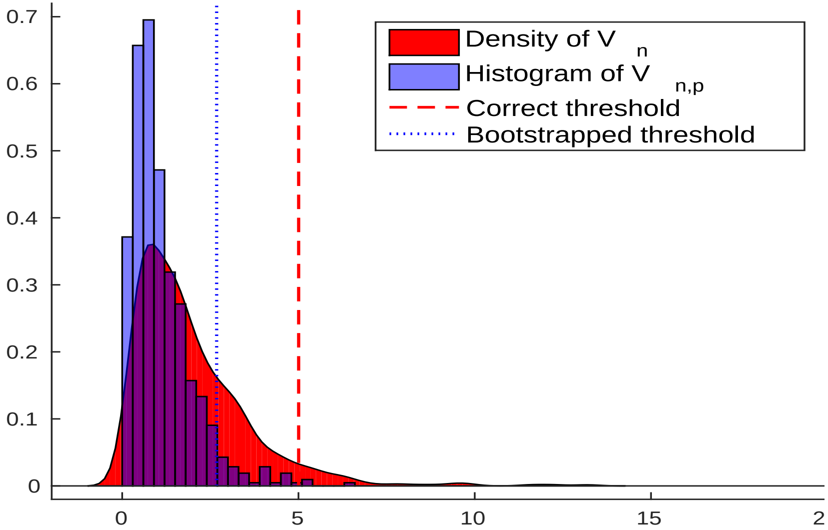
$a = 0.3$

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



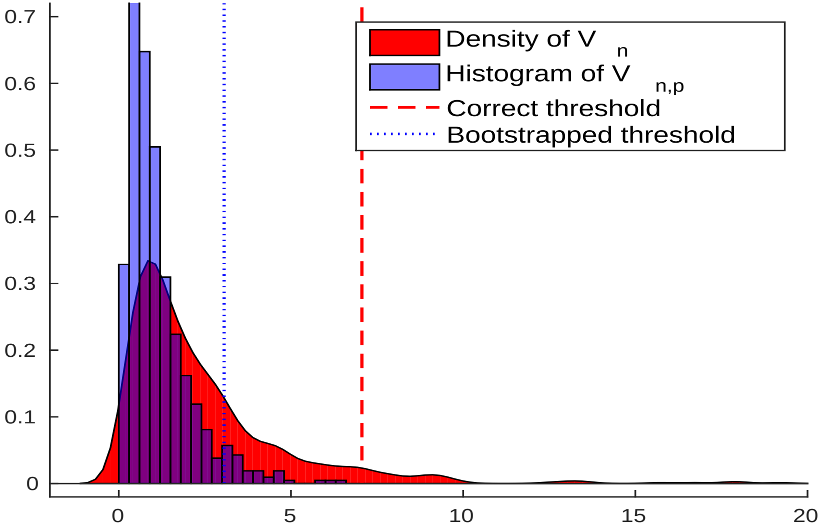
$a = 0.4$

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



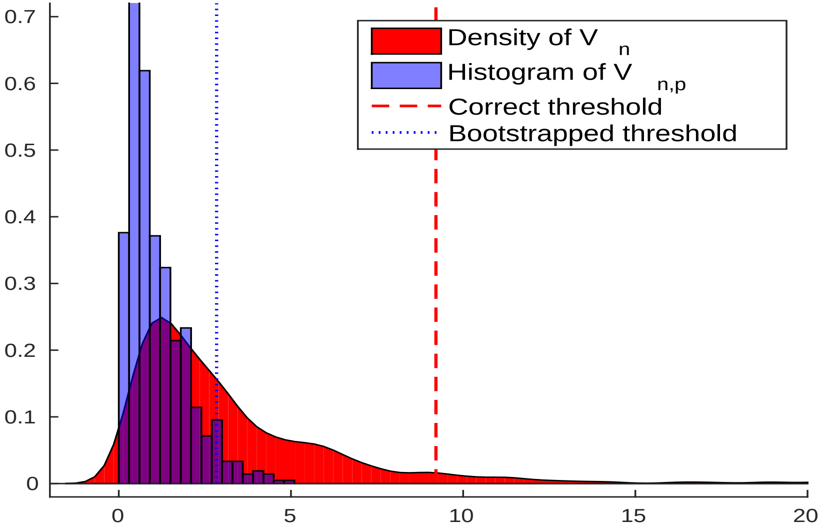
$a = 0.5$

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



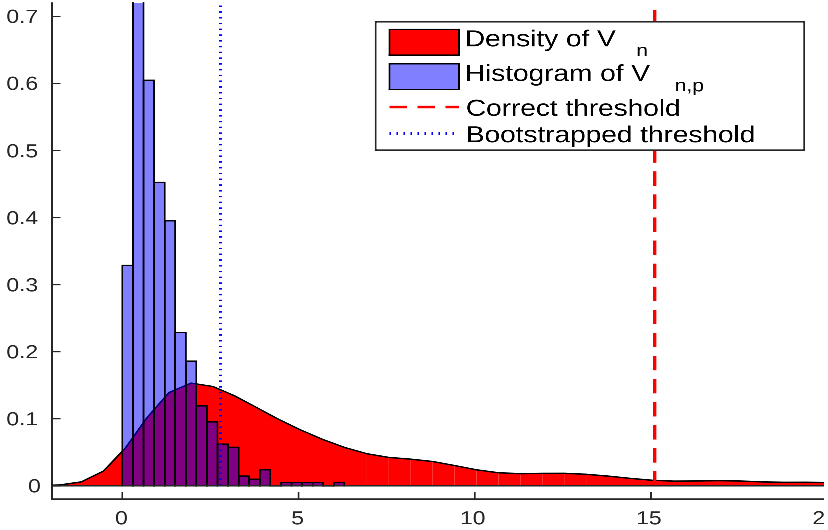
$a = 0.6$

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

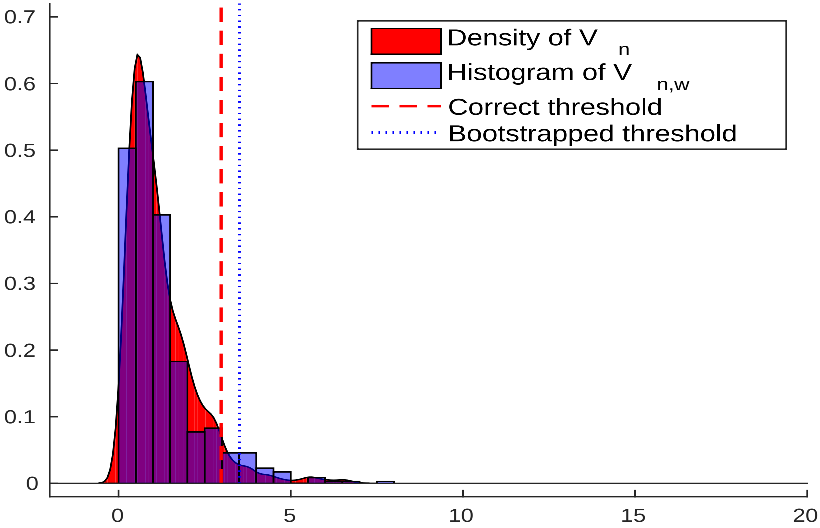
Wild bootstrap process (Leucht and Neumann, 2013):
$W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}} \epsilon_t$ where $W_{0,n}, \epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $\tilde{W}_{t,n} = W_{t,n} - \frac{1}{n}\sum_{j=1}^{n} W_{j,n}$.

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

Theorem (Chwialkowski, S. and Gretton, 2014)

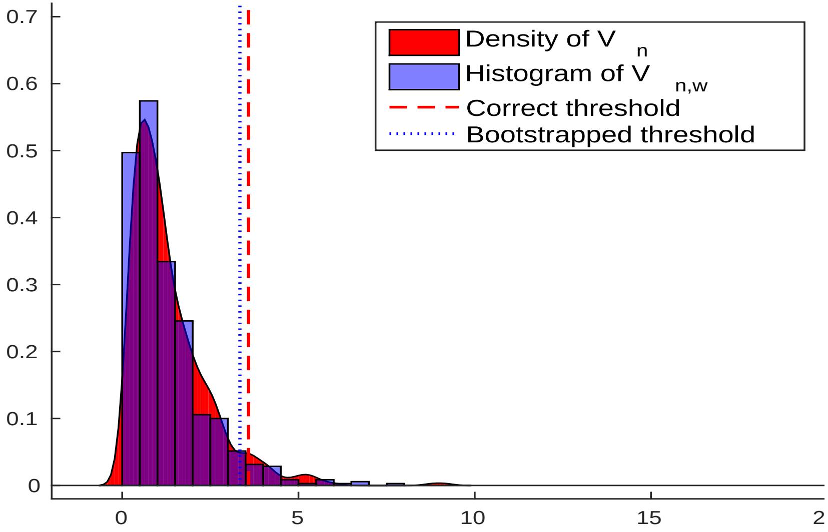
Let k be bounded and Lipschitz continuous, and let $\{X_t\} \sim P$ and $\{Y_t\} \sim Q$ both be τ-dependent with $\tau(r) = O(r^{-6-\epsilon})$, but independent of each other. Then, under $\mathbf{H}_0 : P = Q$, $\varphi\left(\frac{n_x n_y}{n_x+n_y}\widehat{MMD}_k, \frac{n_x n_y}{n_x+n_y}\widehat{MMD}_{k,b}\right)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

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



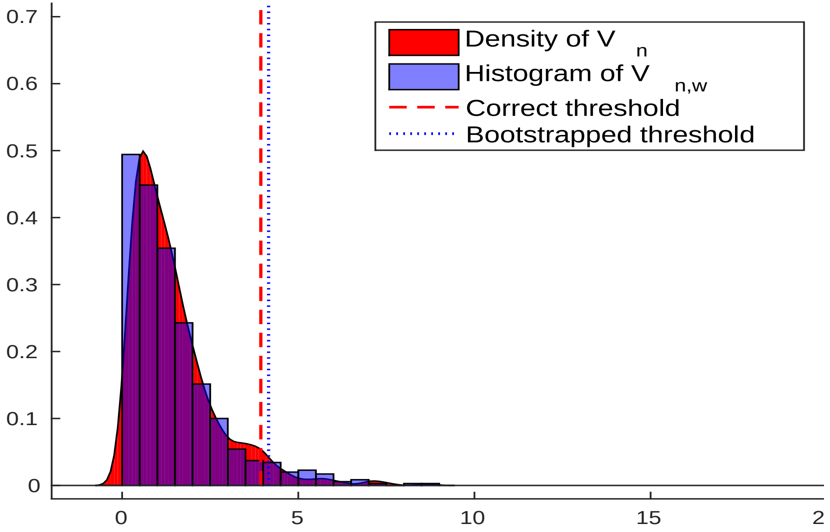
$a = 0.1$

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



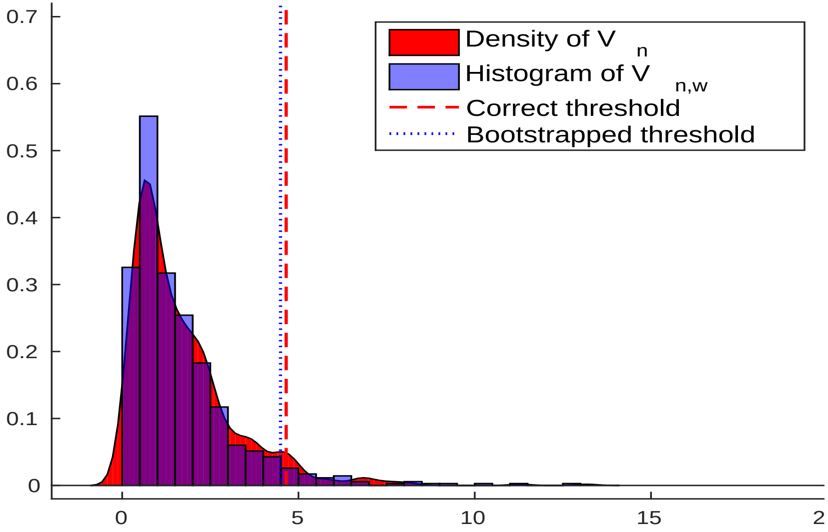
$a = 0.2$

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



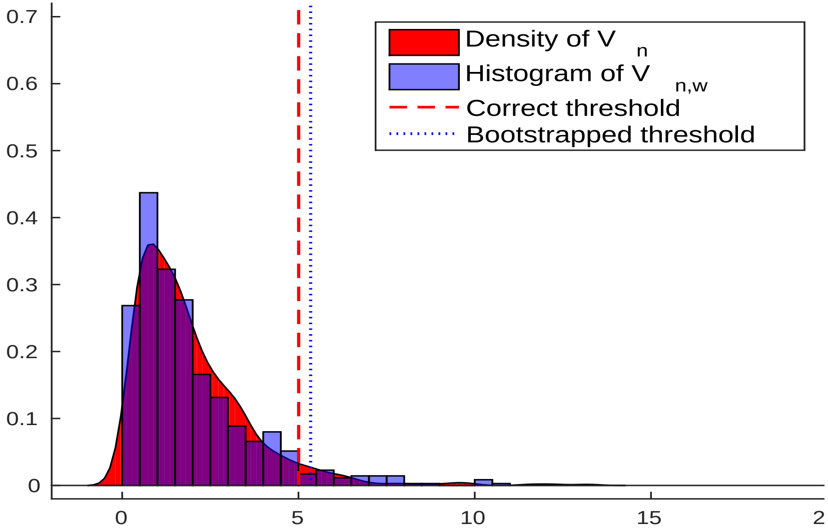
$a = 0.3$

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



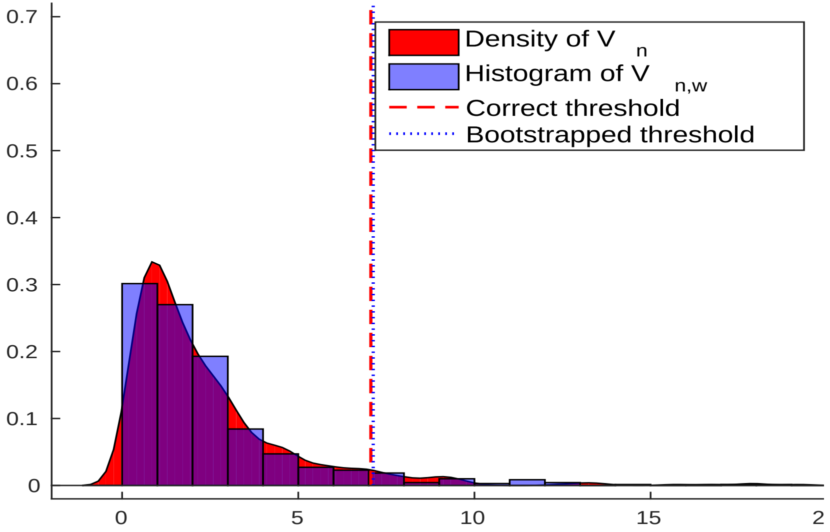
$a = 0.4$

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



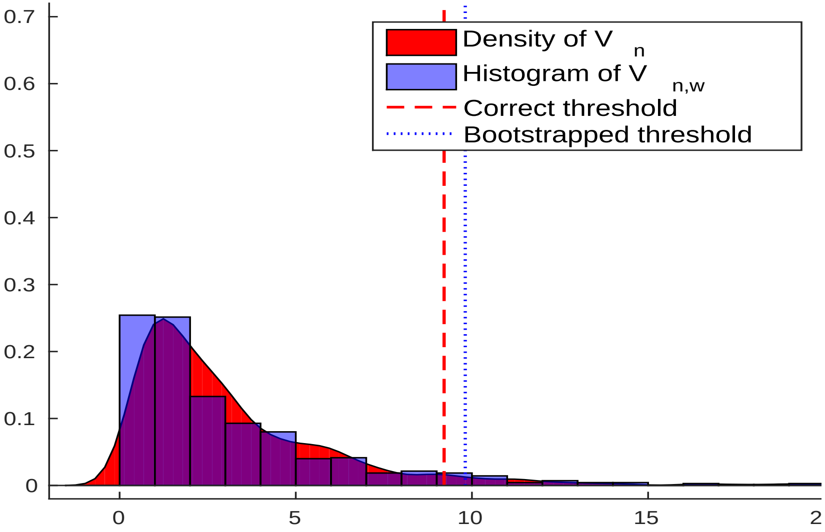
$a = 0.5$

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



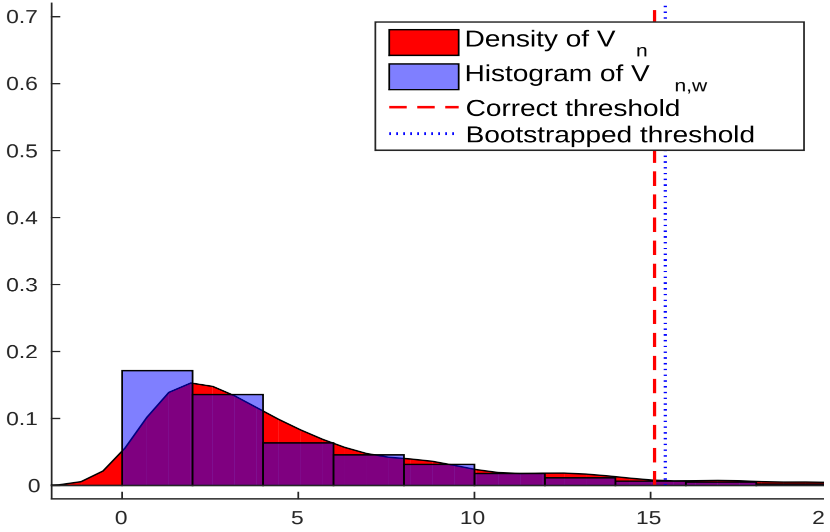
$a = 0.6$

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



$a = 0.7$

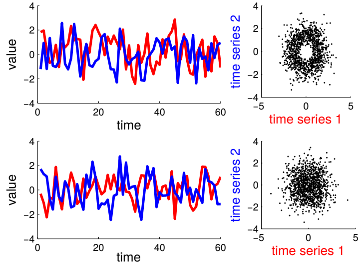
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$

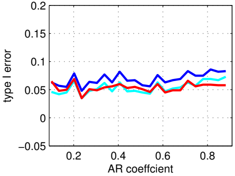


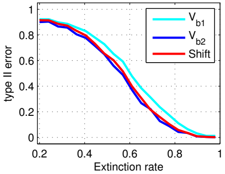
$a = 0.8$

Test calibration for dependent observations

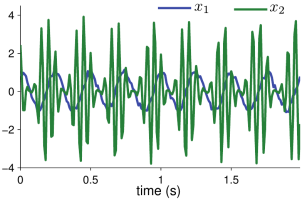
Two-sample test	experiment \ method	perm.	*wild*
MCMC diagnostics	i.i.d. vs i.i.d. ($\mathbf{H_0}$)	.040	.012
	i.i.d. vs Gibbs ($\mathbf{H_0}$)	.528	.052
	Gibbs vs Gibbs ($\mathbf{H_0}$)	.680	.060

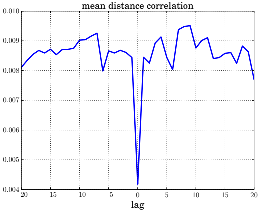


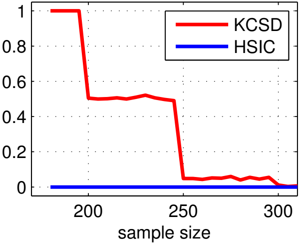




Time Series Coupled at a Lag







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

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

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

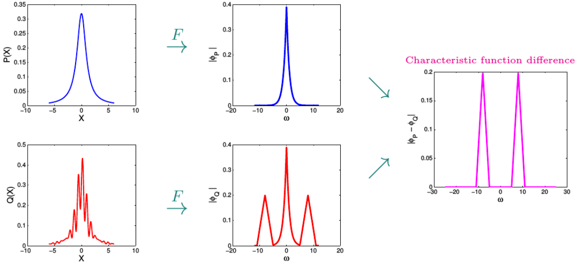
Open questions

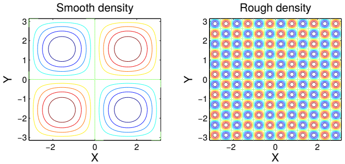
- Interdependent case: how to select parameters of the wild bootstrap / block bootstrap - requires estimating mixing properties of the time series first?
- Large-scale testing: tradeoffs between computation and power
- How to interpret the discovered differences in distributions / discovered dependence? Do we really care about all possible differences between distributions?
- Tuning parameters - can select kernels/hyperparameters to directly optimize relative efficiency of the test, but how does this affect tradeoffs with interdependent data? Sensitive interplay between the kernel hyperparameter and the wild bootstrap parameters
- Multivariate interaction and graphical model selection - approximations?

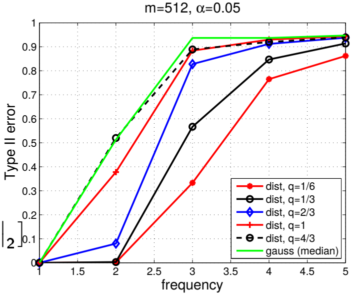
References

- K. Chwialkowski, DS and A. Gretton, **A wild bootstrap for degenerate kernel tests**. *Advances in Neural Information Processing Systems (NIPS)* 27, Dec. 2014.
- DS, 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.
- M. Besserve, N.K. Logothetis and B. Schölkopf, **Statistical analysis of coupled time series with kernel cross-spectral density operators**. *Advances in Neural Information Processing Systems (NIPS)* 26, Dec. 2013.
- A. Leucht and M.H. Neumann, **Dependent wild bootstrap for degenerate U- and V-statistics**. *J. Multivar. Anal.* **117**:257-280, 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.

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.

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 $L_2(P)$.

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

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

Wild Bootstrap

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

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

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

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

ICML Workshop on Large-Scale Kernel Learning

Lille, France, 11 July 2015 (collocated with ICML 2015)

- Foundational algorithmic techniques for large-scale kernel learning: matrix factorization, randomization and approximation, variational inference and sampling, inducing variables, random Fourier features, unifying frameworks
- Interface between kernel methods and deep learning architectures
- Tradeoffs between statistical and computational efficiency in kernel methods
- Stochastic gradient techniques with kernel methods
- Large-scale multiple kernel learning
- Large-scale representation learning with kernels
- Large-scale kernel methods for complex data types beyond perceptual data
- Confirmed speakers: Francis Bach, Neil Lawrence, Russ Salakhutdinov, Marius Kloft, Zaid Harchaoui
- Deadline for Submissions: **Friday, May 1st, 2015**, 23:00 UTC.