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
- highly non-linear associations
- higher-order interactions

\[
X_1 \rightarrow Y \rightarrow X_2
\]
Making Hard Inference Possible

- many dimensions
- low signal-to-noise ratio
- highly non-linear associations
- 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 associations

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

1. Kernel Embeddings and MMD
2. Scaling up Kernel Tests
3. Kernel tests on time series
A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, 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 $\delta_x$ 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 \( x \mapsto (\varphi_1(x), \ldots, \varphi_s(x)) \in \mathbb{R}^s \)

- inner products easily **computed**

- **embedding**: 
  \( P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k \) instead of 
  \( P \mapsto (\mathbb{E}\varphi_1(X), \ldots, \mathbb{E}\varphi_s(X)) \in \mathbb{R}^s \)

- inner products easily **estimated**

\[ \langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X,Y} k(X, Y) \]
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., 
  \( \mathbb{E}_X f(X) = \mathbb{E}_X \langle f, k(\cdot, X) \rangle_{\mathcal{H}_k} = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k} \) \( \forall f \in \mathcal{H}_k \)
Definition

**Kernel metric (MMD)** between $P$ and $Q$:

$$
\text{MMD}_k(P, Q) = \left\| \mathbb{E}_X k(\cdot, X) - \mathbb{E}_Y k(\cdot, Y) \right\|_{\mathcal{H}_k}^2
= \mathbb{E}_{XX'} k(X, X') + \mathbb{E}_{YY'} k(Y, Y') - 2\mathbb{E}_{XY} k(X, Y)
$$
MMD as an integrable probability metric

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

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

- Supremum achieved at the “witness function” \( f = \frac{\mu_k(P) - \mu_k(Q)}{\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}} \).
A polynomial kernel \( k(x, x') = (1 + x^\top x')^s \) 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):

\[
\text{MMD}_k(P_n, P) \to 0 \iff P_n \rightharpoonup P
\]
Nonparametric two-sample tests

- Testing $H_0 : P = Q$ vs. $H_A : P \neq Q$
  based on samples $\{x_i\}_{i=1}^{n_x} \sim P$, $\{y_i\}_{i=1}^{n_y} \sim Q$.

- Test statistic is an estimate of
  $\text{MMD}_k(P, Q) = \mathbb{E}_{XX'}k(X, X') + \mathbb{E}_{YY'}k(Y, Y') - 2\mathbb{E}_{XY}k(X, Y)$:

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

  $- \frac{2}{n_x n_y} \sum_{i,j} k(x_i, y_j)$.

- Degenerate U-statistic: $\frac{1}{\sqrt{n}}$-convergence to MMD under $H_A$,
  $\frac{1}{n}$-convergence to 0 under $H_0$.

- $O(n^2)$ to compute ($n = n_x + n_y$)

Nonparametric independence tests

- $H_0 : X \perp \perp Y$
- $H_A : X \not\perp \perp Y$

Test statistic: $\text{HSIC}(X, Y) = \| \mu_\kappa(\hat{P}_{XY}) - \mu_\kappa(\hat{P}_X \hat{P}_Y) \|_2$

Nonparametric independence tests

- \( H_0 : X \perp \perp Y \Leftrightarrow P_{XY} = P_X P_Y \)
- \( H_A : X \not\perp \not\perp Y \Leftrightarrow P_{XY} \neq P_X P_Y \)

Test statistic:
\[
\text{HSIC}(X, Y) = \left\| \mu_\kappa(\hat{P}_{XY}) - \mu_\kappa(\hat{P}_X \hat{P}_Y) \right\|^2_{\mathcal{H}_\kappa},
\]
with \( \kappa = k \otimes l \)

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...
Kernel Embeddings and MMD

HSIC computation

- **HSIC** measures average similarity between the kernel matrices:
  \[
  \text{HSIC}(X, Y) = \frac{1}{n^2} \langle HKH, HLH \rangle
  \]
- \( H = I - \frac{1}{n} 11^\top \) (centering matrix)
Kernel Embeddings and MMD

**HSIC computation**

\[ k(x_i, x_j) \rightarrow K = \]

\[ \ell(x_i, x_j) \rightarrow L = \]

- **HSIC** measures *average similarity between the kernel matrices*:
  \[ \text{HSIC}(X, Y) = \frac{1}{n^2} \langle HKH, HLLH \rangle \]
  \[ H = I - \frac{1}{n}11^\top \text{ (centering matrix)} \]

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

\[ \| \mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y) \|_{\mathcal{H}_\kappa} = \sup_f \left[ \mathbb{E}_{X,Y} f(X, Y) - \mathbb{E}_X \mathbb{E}_Y f(X, Y) \right] \]

- witness lies in the unit ball of \( \mathcal{H}_\kappa = \mathcal{H}_k \otimes \mathcal{H}_l \), the RKHS of functions on \( X \times Y \)
Three-variable interaction and V-structure discovery

\[ \Delta_L P = P_{XYZ} - P_{XYP} - P_{XZY} - P_{XYZ} + 2P_{XYPZ} \]

V-structure discovery: Dataset A

Outline

1. Kernel Embeddings and MMD

2. Scaling up Kernel Tests

3. Kernel tests on time series
Test threshold

- under $H_0 : P = Q$:
  $$\frac{n_x n_y}{n_x + n_y} \text{MMD}_k \sim \sum_{r=1}^{\infty} \lambda_r (Z_r^2 - 1), \quad \{Z_r\} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$$

  - $\{\lambda_r\}$ depend on both $k$ and $P$: eigenvalues of $T : L_2 \rightarrow L_2$, 
    $$(Tf)(x) \mapsto \int f(x') \tilde{k}(x, x') dP(x').$$

- expensive threshold computation:
  - Estimate leading $\lambda_r$'s (eigendecomposition of the kernel matrix): $O(n^3)$
  - Permutation test: $\# \text{shuffles} \times O(n^2)$
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

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

- Complexity: \( O(n^2) \).
Limited time, unlimited data

- Process mini-batches of size $B = B_x + B_y$ at a time:
  $$\hat{\eta}_k = \frac{B}{n} \sum_{b=1}^{n/B} \hat{\text{MMD}}_{k,b}$$
- Complexity:
  $$O(nB) = \frac{n}{B} \times O(B^2)$$
- Provided $B/n \rightarrow 0$:
  - $\frac{1}{\sqrt{n}}$-convergence to MMD under $H_A$,
  - $\frac{1}{\sqrt{nB}}$-convergence to 0 under $H_0$.

Null distribution

\[
\frac{n_x n_y}{(n_x + n_y)^{3/2}} \sqrt{B} \hat{\eta}_k \overset{\text{d}}{\sim} \mathcal{N}(0, \sigma_k^2) \text{ under } H_0.
\]

- \(\sigma_k^2\) (depends on \(k\) and \(P\)) can be unbiasedly estimated on each block \(b\) in \(O(B^2)\) time:

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

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

- Alternatively, track empirical variance of \(\left\{\hat{\text{MMD}}_{k,b}\right\}_{b=1}^{n/B}\).
- No need for permutation testing.
### Full statistic vs. mini-batch statistic

<table>
<thead>
<tr>
<th></th>
<th>$U$-statistic</th>
<th>mini-batch</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>$O(n^2)$</td>
<td>$O(nB)$</td>
</tr>
<tr>
<td>storage</td>
<td>$O(n^2)$</td>
<td>$O(B^2)$</td>
</tr>
<tr>
<td>null distribution</td>
<td>infinite sum of chi-squares</td>
<td>normal</td>
</tr>
<tr>
<td>computing p-value</td>
<td>$O(n^3)$ or #shuffles $\times O(n^2)$</td>
<td>$O(nB)$</td>
</tr>
<tr>
<td>$H_0$-convergence rate</td>
<td>$1/n$</td>
<td>$1/\sqrt{nB}$</td>
</tr>
</tbody>
</table>
Asymptotic efficiency criterion


**Proposition**

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

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

minimizes the asymptotic (Hodges-Lehmann) relative efficiency on $\mathcal{K}$. 

We only have estimates of $\eta_k$ and $\sigma_k$!

Will the kernel optimization using plug-in estimates be consistent?

Over what families of kernels can we perform such optimization efficiently?
Asymptotic efficiency criterion


**Proposition**

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

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

minimizes the asymptotic (Hodges-Lehmann) relative efficiency on $\mathcal{K}$.

- We only have estimates of $\eta_k$ and $\sigma_k$!
- Will the kernel optimization using plug-in estimates be consistent?
- Over what families of kernels can we perform such optimization efficiently?
Asymptotic efficiency criterion


**Proposition**

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

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

minimizes the asymptotic (Hodges-Lehmann) relative efficiency on $\mathcal{K}$.

- We only have estimates of $\eta_k$ and $\sigma_k$!
- Will the kernel optimization using plug-in estimates 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: $3 \times 3$ 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.

<table>
<thead>
<tr>
<th></th>
<th>$m$ per trial</th>
<th>Type II error</th>
<th>Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic</td>
<td>5,000</td>
<td>[0.7996, 0.8516]</td>
<td>820</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>[0.5161, 0.6175]</td>
<td>367</td>
</tr>
<tr>
<td></td>
<td>&gt; 10,000</td>
<td>Buy more RAM!</td>
<td></td>
</tr>
</tbody>
</table>

Buy more RAM!
Gaussian blobs (2)

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

<table>
<thead>
<tr>
<th></th>
<th>$m$ per trial</th>
<th>Type II error</th>
<th>Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quadratic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5,000</td>
<td>$[0.7996, 0.8516]$</td>
<td>820</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>$[0.5161, 0.6175]$</td>
<td>367</td>
</tr>
<tr>
<td></td>
<td>$&gt; 10,000$</td>
<td>Buy more RAM!</td>
<td></td>
</tr>
<tr>
<td><strong>Linear</strong></td>
<td>$\sim$ 100,000,000</td>
<td>$[0.2250, 0.3049]$</td>
<td>468</td>
</tr>
<tr>
<td></td>
<td>$\sim$ 200,000,000</td>
<td>$[0.1873, 0.2829]$</td>
<td>302</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$\sim$ 500,000,000</td>
<td><strong>0.0270 $\pm$ 0.0302</strong></td>
<td>111</td>
</tr>
</tbody>
</table>
Figure: $m = 10,000$; family generated by gaussian kernels with bandwiths $\{2^{-5}, \ldots, 2^{15}\}$. 

**Note:** This figure illustrates the Type II error as a function of the ratio between eigenvalues for different kernel tests. The legend indicates the different tests used: Opt, $L_2$-MaxMMD, MaxMMD, and Median.
Hard-to-detect differences: UCI HIGGS


- Benchmark dataset for distinguishing a signature of Higgs boson vs. background
- Joint distributions of the azimuthal angular momenta $\varphi$ for four particle jets: low-signal, low-level features
- Do joint angular momenta carry any discriminating information?
Hard-to-detect differences: UCI HIGGS


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

<table>
<thead>
<tr>
<th>sample size:</th>
<th>1e4</th>
<th>5e4</th>
<th>1e5</th>
<th>5e5</th>
<th>1e6</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value (gauss-med):</td>
<td>.757</td>
<td>.217</td>
<td>.475</td>
<td>.391</td>
<td>.074</td>
</tr>
</tbody>
</table>
Hard-to-detect differences: UCI HIGGS


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

<table>
<thead>
<tr>
<th>sample size:</th>
<th>1e4</th>
<th>5e4</th>
<th>1e5</th>
<th>5e5</th>
<th>1e6</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value (gauss-med):</td>
<td>.757</td>
<td>.217</td>
<td>.475</td>
<td>.391</td>
<td>.074</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>train/test size:</th>
<th>2e3/8e3</th>
<th>1e4/4e4</th>
<th>2e4/8e4</th>
<th>1e5/4e5</th>
<th>2e5/8e5</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value (gauss-opt):</td>
<td>.139</td>
<td>.476</td>
<td>.035</td>
<td>$6.12e-5$</td>
<td>$1.02e-18$</td>
</tr>
</tbody>
</table>
Experiment: Independence Test ($\sum \text{sign } \prod$)

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

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

$$MMD_k(P, Q) = \| \mathbb{E}_X k(\cdot, X) - \mathbb{E}_Y k(\cdot, Y) \|_{\mathcal{H}_k}^2$$

$$= \mathbb{E}_{XX'} k(X, X') + \mathbb{E}_{YY'} k(Y, Y') - 2 \mathbb{E}_{X,Y} k(X, Y)$$
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2} \epsilon_t$
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Permutation test on AR(1): \( X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t \)
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2} \epsilon_t$
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Permutation test on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Wild Bootstrap

Wild bootstrap process (Leucht and Neumann, 2013):

\[ W_{t,n} = e^{-1/ln} W_{t-1,n} + \sqrt{1 - e^{-2/ln}} \epsilon_t \]

where \( W_{0,n}, \epsilon_1, \ldots, \epsilon_n \) \( i.i.d. \) \( \mathcal{N}(0,1) \), and

\[ \tilde{W}_{t,n} = W_{t,n} - \frac{1}{n} \sum_{j=1}^{n} W_{j,n} \].

\[ \overline{\text{MMD}}_{k,wb} := \frac{1}{n_x^2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} \tilde{W}_{i,n_x}^{(x)} \tilde{W}_{j,n_x}^{(x)} k(x_i, x_j) - \frac{1}{n_x^2} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} \tilde{W}_{i,n_y}^{(y)} \tilde{W}_{j,n_y}^{(y)} k(y_i, y_j) \]

\[ - \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \tilde{W}_{i,n_x}^{(x)} \tilde{W}_{j,n_y}^{(y)} k(x_i, y_j) \].

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 \( \tau \)-dependent with \( \tau(r) = O(r^{-6-\epsilon}) \), but independent of each other.

Then, under \( H_0 : P = Q \),

\[ \varphi \left( \frac{n_x n_y}{n_x + n_y} \overline{\text{MMD}}_k, \frac{n_x n_y}{n_x + n_y} \overline{\text{MMD}}_{k,b} \right) \overset{P}{\to} 0 \quad \text{as } n_x, n_y \to \infty, \]

where \( \varphi \) is the Prokhorov metric.
**Mercer’s Expansion**

For a compact metric space $\mathcal{X}$, and a continuous 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 $L_2(P)$.

$$\mathcal{H}_k \ni k(\cdot, x) \iff \left\{ \sqrt{\lambda_r} \Phi_r(x) \right\} \in \ell_2$$

$$\mathcal{H}_k \ni \mu_k(P) \iff \left\{ \sqrt{\lambda_r} \mathbb{E}\Phi_r(X) \right\} \in \ell_2$$

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

- \( \rho_x = n_x/n, \rho_y = n_y/n \)

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

\[
\rho_x \rho_y \sqrt{n} \text{MMD}^k = \sum_{r=1}^{\infty} \lambda_r \left( \sqrt{\rho_y} \sum_{t=1}^{n_x} \frac{\Phi_r(X_t)}{\sqrt{n_x}} - \sqrt{\rho_x} \sum_{t=1}^{n_y} \frac{\Phi_r(Y_t)}{\sqrt{n_y}} \right)^2
\]

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

- \( \mathbb{E} [ \Phi_r(X_1) W_{1,n} \Phi_s(X_t) W_{t,n} ] = \mathbb{E} [ \Phi_r(X_1) \Phi_s(X_t) ] \zeta \left( \frac{|t-1|}{\ell_n} \right) \to \mathbb{E} [ \Phi_r(X_1) \Phi_s(X_t) ], \forall t, r, s \) provided dependence between \( X_1 \) and \( X_t \) “disappears fast enough” (a \( \tau \)-mixing condition).
Wild Bootstrap on AR(1): \( X_{t+1} = aX_t + \sqrt{1-a^2}\epsilon_t \)
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$

$a = 0.3$

Density of $V_n$

Histogram of $V_{n,w}$

Correct threshold

Bootstrapped threshold
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Wild Bootstrap on AR(1): \( X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t \)
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2} \epsilon_t$
Wild Bootstrap on AR(1): $X_{t+1} = aX_t + \sqrt{1 - a^2}\epsilon_t$
Wild Bootstrap on AR(1): \( X_{t+1} = aX_t + \sqrt{1 - a^2} \epsilon_t \)
Test calibration for dependent observations

<table>
<thead>
<tr>
<th>Two-sample test</th>
<th>experiment \ method</th>
<th>perm.</th>
<th>wild</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCMC diagnostics</td>
<td>i.i.d. vs i.i.d. ($H_0$)</td>
<td>.040</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>i.i.d. vs Gibbs ($H_0$)</td>
<td>.528</td>
<td>.052</td>
</tr>
<tr>
<td></td>
<td>Gibbs vs Gibbs ($H_0$)</td>
<td>.680</td>
<td>.060</td>
</tr>
</tbody>
</table>
Time Series Coupled at a Lag

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

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

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

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 ($\tau$-mixing), applicable to both two-sample and independence tests.
- Applications: MCMC diagnostics, time series dependence across multiple lags.
Kernel tests on time series

References

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 $\chi^2$)
- 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


\[ \nu^2(X, Y) = \mathbb{E}_{XY} \mathbb{E}_{X'Y'} \|X - X'\|_2 \|Y - Y'\|_2 \\
+ \mathbb{E}_X \mathbb{E}_{X'} \|X - X'\|_2 \mathbb{E}_Y \mathbb{E}_{Y'} \|Y - Y'\|_2 \\
- 2 \mathbb{E}_{XY} \left[ \mathbb{E}_{X'} \|X - X'\|_2 \mathbb{E}_{Y'} \|Y - Y'\|_2 \right] \\
- \frac{d}{dCov} \left( \mathbb{E}_X \|X - X'\|_2 \mathbb{E}_Y \|Y - Y'\|_2 \right) \\
\]