The effect of kernel choice on RKHS based statistical tests

Bharath K. Sriperumbudur,[1] Arthur Gretton,[2]
Kenji Fukumizu,[3] & Bernhard Schölkopf[2]

[1]University of California, San Diego; [2]MPI for Biological Cybernetics, Tübingen;
[3]Institute for Statistical Mathematics, Tokyo.

NIPS 2007 Workshop

Representations and Inference on Probability Distributions

Two-Sample Problem

- Given:
  - m samples $\mathbf{X} := \{x_1, \ldots, x_m\}$ drawn i.i.d. from $\mathbf{P}$.
  - n samples $\mathbf{Y} := \{y_1, \ldots, y_n\}$ drawn i.i.d. from $\mathbf{Q}$.
- Determine: are $\mathbf{P}$ and $\mathbf{Q}$ different.
- Applications:
  - Microarray data aggregation
  - Speaker/author identification
  - Schema matching
- Issues: To deal with
  - High dimensionality
  - Low sample size
  - Structured domains (strings and graphs)

Maximum Mean Discrepancy (MMD)

Lemma ([Dudley, 2002])

Let $(\mathcal{X}, d)$ be a separable metric space, and let $\mathbf{P}$, $\mathbf{Q}$ be two Borel probability measures defined on $\mathcal{X}$. Then $\mathbf{P} = \mathbf{Q}$ if and only if $\mathbb{E}_{\mathbf{P}}[f(x)] = \mathbb{E}_{\mathbf{Q}}[f(x)], \forall f \in \mathscr{C}(\mathcal{X})$, where $\mathscr{C}(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.

- Test statistic: [Gretton et al., 2007]

$$MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] := \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbf{P}}[f(x)] - \mathbb{E}_{\mathbf{Q}}[f(y)] \right). \tag{1}$$

  for some function class $\mathcal{F}$.

- $\mathcal{F} = \mathscr{C}(\mathcal{X})$: $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0 \Leftrightarrow \mathbf{P} = \mathbf{Q}$.
- Is there any other function class $\mathcal{F}$ apart from $\mathscr{C}(\mathcal{X})$ for which $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0 \Leftrightarrow \mathbf{P} = \mathbf{Q}$?

Maximum Mean Discrepancy (MMD)

Theorem ([Gretton et al., 2007])

Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathcal{X}$, with associated kernel $k(.,.)$. Then $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0$ if and only if $\mathbf{P} = \mathbf{Q}$.

- Test statistic:
  - $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = \left\| \mathbb{E}_{\mathbf{P}}[k(.,x)] - \mathbb{E}_{\mathbf{Q}}[k(.,y)] \right\|_{\mathcal{H}}$
  - $\widehat{MMD}[\mathcal{F}, m, n] = \left\| \frac{1}{n} \sum_{i=1}^{n} k(.,x_i) - \frac{1}{m} \sum_{i=1}^{m} k(.,y_i) \right\|_{\mathcal{H}}$
- $\mathbf{P} = \mathbf{Q}$, $m = n$, $k(x,x) \leq K < \infty$:
  - Consistency: $\widehat{MMD}[\mathcal{F}, n, n] = O\left(\frac{1}{\sqrt{n}}\right)$
- Experimentally, the method is shown to work well on small sample sizes, high dimensional data and is even applicable to data from structured domains.

When will the method fail?

- $k(.,.)$ induces $\mathcal{H}$. So, the method is as good as the kernel.
- Universal RKHS: [Steinwart, 2002]
  - When $\mathcal{X}$ is compact, $\mathcal{H}$ is dense in $\mathscr{C}(\mathcal{X})$ with respect to the L_∞ norm.
  - Universal kernels: Gaussian, Laplacian.
- Questions:
  - Are there non-universal kernels for which $\exists \mathbf{P} \neq \mathbf{Q}$ such that $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0$?
  - For what class of probability distributions, can a non-universal kernel behave as: $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0 \Leftrightarrow \mathbf{P} = \mathbf{Q}$?
  - Are there non-universal kernels for which $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0 \Leftrightarrow \mathbf{P} = \mathbf{Q}, \forall\, \mathbf{P}, \mathbf{Q}$?
- New formulation is needed to answer these questions.

Background & Notation

Assumption ①: $\mathcal{X} \subseteq \mathbb{R}^d$. $k(.,.)$ is translation-invariant, i.e., $k(x,y) = \psi(x-y)$, where $\psi \in \mathscr{C}(\mathbb{R}^d)$ is a positive definite function.

- By Bochner's theorem, $\psi(x) = \int_{\mathbb{R}^d} \exp(-j\langle \omega, x \rangle)\, d\Lambda(\omega)$, $x \in \mathbb{R}^d$, where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.
- $\Psi(\omega) := \frac{d\Lambda}{d\omega}$ is the distributional derivative of Λ.
- Characteristic function of $\mathbf{P}$:
  $\phi_{\mathbf{P}}(\omega) := \int_{\mathbb{R}^d} \exp(j\langle \omega, x \rangle)\, d\mathbf{P}(x)$, $\omega \in \mathbb{R}^d$.
- $p(x) := \frac{d\mathbf{P}}{dx}$ is the distributional derivative of $\mathbf{P}$. Similarly q is the distributional derivative of $\mathbf{Q}$.

New Formulation of MMD

Theorem

Let $\mathcal{F}$ be a unit ball in a RKHS $\mathcal{H}$ (not necessarily universal), defined on $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\phi_{\mathbf{P}}$ and $\phi_{\mathbf{Q}}$ be the characteristic functions corresponding to $\mathbf{P}$ and $\mathbf{Q}$ respectively. Suppose $k(.,.)$ satisfies ①. Then

$$MMD[\mathcal{F},\mathbf{P},\mathbf{Q}] = \left\|\mathbb{F}^{-1}\left[\Psi(\phi^*_{\mathbf{P}} - \phi^*_{\mathbf{Q}})\right]\right\|_{\mathcal{H}}, \tag{2}$$

where $\mathbb{F}^{-1}$ is the Fourier inverse and * is the complex conjugation.

The above formulation is used to study the behavior of MMD, more specifically the case of $MMD[\mathcal{F},\mathbf{P},\mathbf{Q}] = 0$.

Characteristic Kernel

Definition (Characteristic kernel)

A positive-definite kernel is a characteristic kernel for a class, $\mathcal{D}$ of probability measures on $\mathbb{R}^d$ if $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0 \Leftrightarrow \mathbf{P} = \mathbf{Q}$ for all $\mathbf{P}, \mathbf{Q} \in \mathcal{D}$.

Remark: Universal kernels on a compact subset of $\mathbb{R}^d$ are characteristic kernels for any $\mathbf{P}, \mathbf{Q}$.

Example (Non-characteristic kernel)

Let $\psi(x) = 1, \forall\, x \in \mathbb{R}^d$. Then $\Psi(\omega) = (2\pi)^d \delta(\omega)$, i.e., $\Psi(\omega) = 0,\ \omega \in \mathbb{R}^d \backslash \{0\}$. Therefore, $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0,\ \forall\, \mathbf{P}, \mathbf{Q}$.

Question: Are there interesting kernels for which $\exists \mathbf{P} \neq \mathbf{Q}$ such that $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0$?

Main Result

Theorem

Let $\mathcal{F}$ be a unit ball in a RKHS $\mathcal{H}$ defined on $\mathcal{X} \subseteq \mathbb{R}^d$. Suppose that $k(.,.)$ satisfies ① and $supp(\Psi) \subseteq \mathbb{R}^d$. Let $\mathbf{P}, \mathbf{Q}$ be probability distributions on $\mathbb{R}^d$ such that $\mathbf{P} \neq \mathbf{Q}$. Then $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0$ if and only if there exists a tempered distribution $\theta : \mathscr{S} \to \mathbb{C}$ that satisfies the following conditions:

(i) $p - q = \mathbb{F}^{-1}\theta$

(ii) $\theta\Psi = 0$

where $\mathscr{S}$ is the Schwartz space of rapidly decaying functions.

Remarks

- Dependence on the kernel: through $\mathrm{supp}(\Psi)$.
- Three cases: Suppose $\mathcal{X} \subseteq \mathbb{R}$.
  1. $\{\omega : \Psi(\omega) = 0\}$ is empty
  2. $\{\omega : \Psi(\omega) = 0\}$ is non-empty but countable
  3. $\{\omega : \Psi(\omega) = 0\}$ is uncountable
- The following proposition settles the case when $\{\omega : \Psi(\omega) = 0\}$ is empty.

Proposition

Let ψ be such that $\Psi(\omega) > 0, \forall \omega \in \mathbb{R}^d$. Then ψ is a characteristic kernel for any $\mathcal{D}$.

Example: Gaussian and Laplacian kernels.

$\{\omega : \Psi(\omega) = 0\}$ is non-empty but countable

Proposition

Let ψ be such that supp$(\Psi) = \mathbb{R}^d$. Then ψ is a characteristic kernel for any $\mathcal{D}$.

Example: B_{2n+1}-spline kernels.

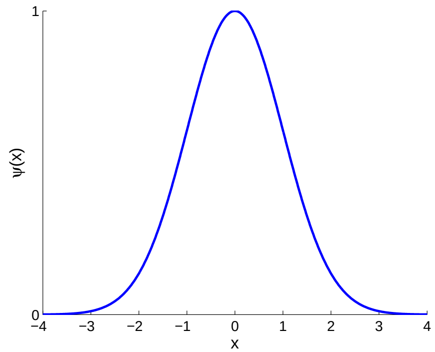
Corollary

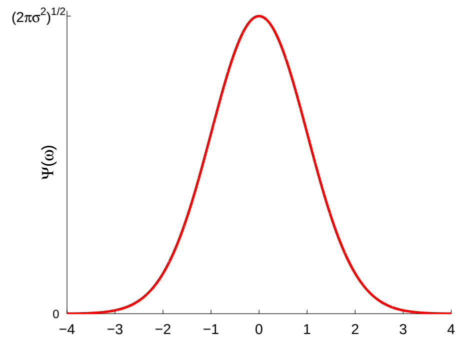
Let ψ be compactly supported on $\mathbb{R}^d$. Then ψ is a characteristic kernel for any $\mathcal{D}$.

Advantage: Compactly supported kernels are computationally advantageous compared with non-compact kernels such as the Gaussian and Laplacian.

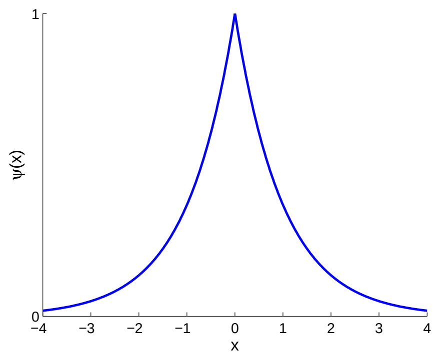
Examples of characteristic kernels (for any $\mathcal{D}$)

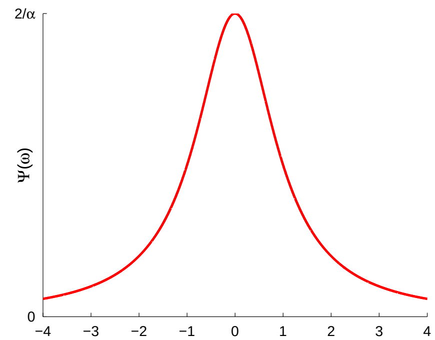
- Gaussian kernel:



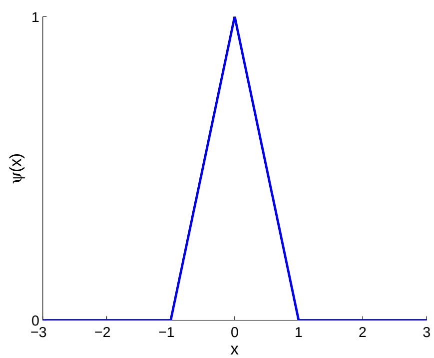


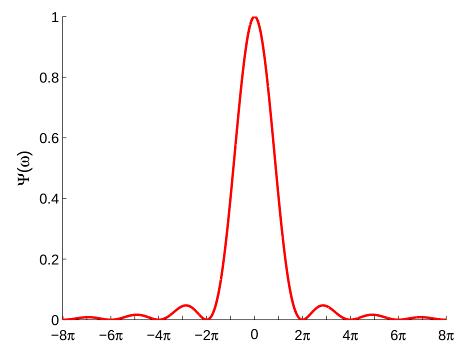
- Laplacian kernel:





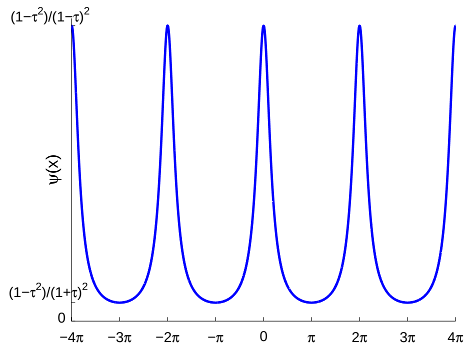
- B_1-spline kernel:

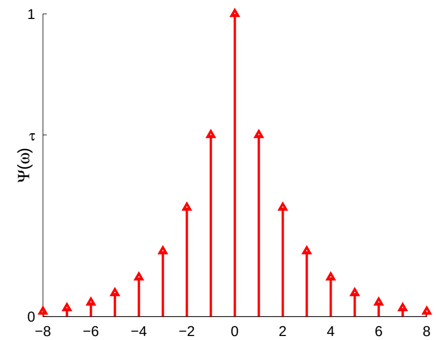




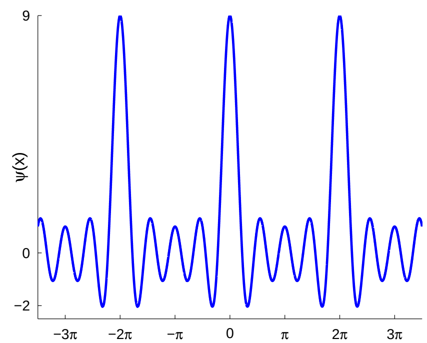
$\{\omega : \Psi(\omega) = 0\}$ is uncountable : Examples

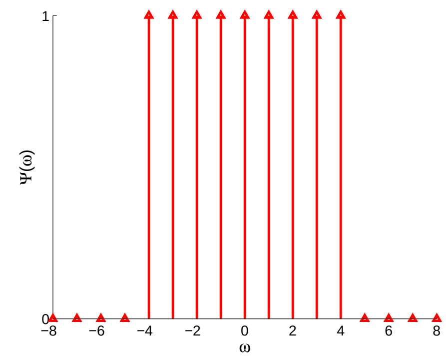
- Poisson kernel:



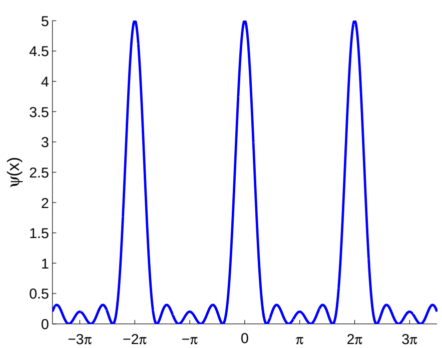


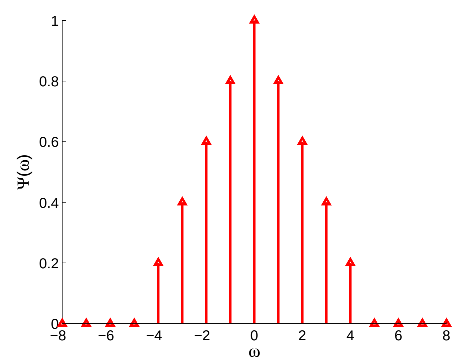
- Dirichlet kernel:





- Féjer kernel:





$\{\omega : \Psi(\omega) = 0\}$ is uncountable

Proposition

Let $\mathcal{D}$ be the class of discrete probability measures defined on $\mathcal{X}$. Then $\exists \mathbf{P} \neq \mathbf{Q}$, $\mathbf{P}, \mathbf{Q} \in \mathcal{D}$ such that $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0$ if and only if the following conditions hold:

(i) *ψ is τ-periodic on $\mathbb{R}^d$, i.e.,*
$\psi(x) = \psi(x + \gamma \circ \tau)$, $\gamma \in \mathbb{Z}^d$, $0 \prec \tau \prec \infty$.

(ii) $\mathcal{X} = \{x_1, x_2, \ldots\}$, $x_n \in \{l \circ \tau : l \in \mathbb{Z}^d\}$, $\forall\, n$, *where $\circ$ represents the Hadamard multiplication.*

- This is a very limited case for the test to fail.
- Every aperiodic kernel is a characteristic kernel for the class of discrete probability measures.
  - Example: $\psi(x) = \frac{sin(Mx)}{\pi x}$ with $\Psi(\omega) = \mathbf{1}_{[-M,M]}(\omega)$, $x, \omega \in \mathbb{R}$.

$\{\omega : \Psi(\omega) = 0\}$ is uncountable

Proposition

Let $\mathcal{D}$ be the class of non-discrete probability measures that are compactly supported on $\mathbb{R}^d$. Suppose ψ be such that supp$(\Psi) \subset \mathbb{R}^d$. Then ψ is a characteristic kernel for $\mathcal{D}$.

Proof idea: Based on the following result (a corollary of Paley-Wiener theorem).

Lemma ([Mallat, 1998])

If $g \neq 0$ has a compact support then its Fourier transform, $G(\omega)$ cannot be zero on a whole interval. Similarly, if $G \neq 0$ has a compact support then $g(x)$ cannot be zero on a whole interval.

$\{\omega : \Psi(\omega) = 0\}$ is uncountable

- Non-discrete probability measures with non-compact support:
  - Does there exist $\theta \neq 0$ satisfying the conditions in the main result?
  - The following result due to Paley & Wiener can be used to address this issue. [Strichartz, 2003]

Theorem (Paley-Wiener)

Let g be a C^∞ function supported in $[-M, M]$. Then $G(\omega + j\sigma)$ is an entire function of exponential type M, i.e. $\exists C$ such that

$$|G(\omega + j\sigma)| \leq C \exp(M|\sigma|), \tag{3}$$

and $G(\omega)$ is rapidly decreasing, i.e., $\exists\, c_n$ such that

$$|G(\omega)| \leq c_n(1 + |\omega|)^{-n}, \forall\, n \in \mathbb{N}. \tag{4}$$

In addition, the converse also holds.

$\{\omega : \Psi(\omega) = 0\}$ is uncountable

- Existence of $g \in C^\infty$ supported in $[-M, M]$:
  $g_{M,\omega_0}(\omega) = \mathbf{1}_{(-M,M)}(\omega - \omega_0) \exp\left(-\frac{M^2}{M^2-(\omega-\omega_0)^2}\right)$.
- Choose $\theta(\omega) = g_{M,\omega_0}(\omega)$ for some M, ω_0 so that $\theta(\omega)$ satisfies the conditions in the main result.
- $\mathbb{F}^{-1}\theta$ is a rapidly decaying function.

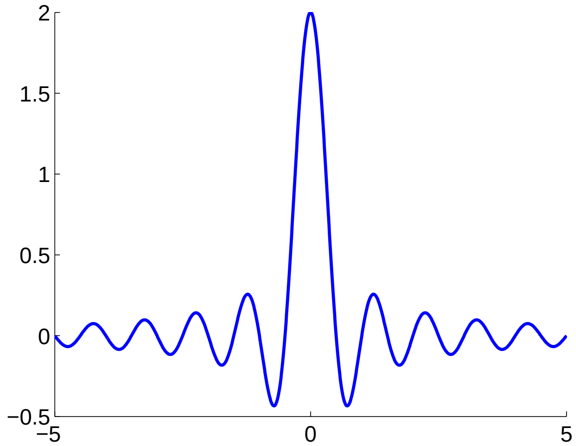
With the above construction, we have the following result:

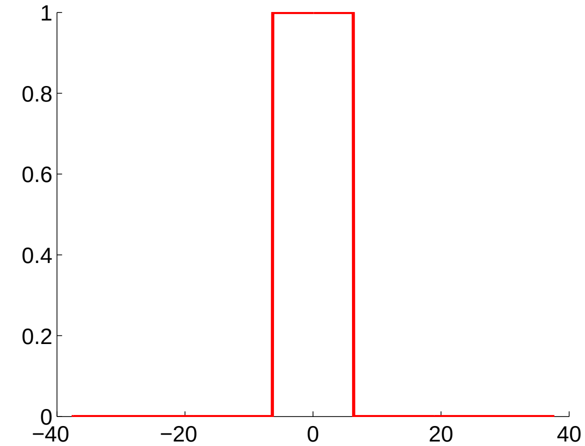
Proposition

Let $\mathcal{D}$ be the class of non-discrete probability measures that are non-compactly supported on $\mathbb{R}^d$. Suppose ψ be such that supp$(\Psi) \subset \mathbb{R}^d$. Then $\exists \mathbf{P} \neq \mathbf{Q}$ such that $MMD[\mathcal{F}, \mathbf{P}, \mathbf{Q}] = 0$.

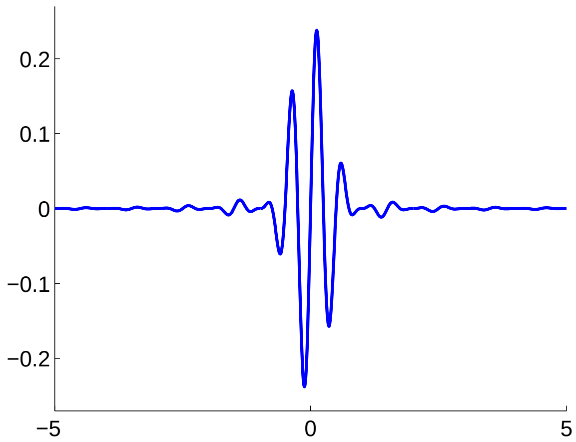
Example

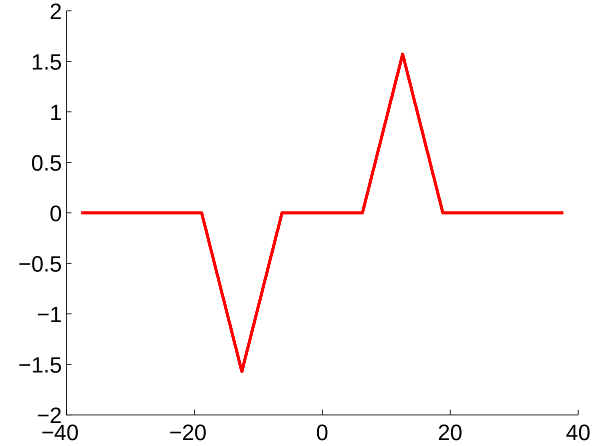
- $\psi(x) = \frac{sin(Mx/2)}{\pi x} = \frac{M}{\pi} sinc\left(\frac{Mx}{\pi}\right)$; $\Psi(\omega) = \mathbf{1}_{[-M,M]}(\omega)$.



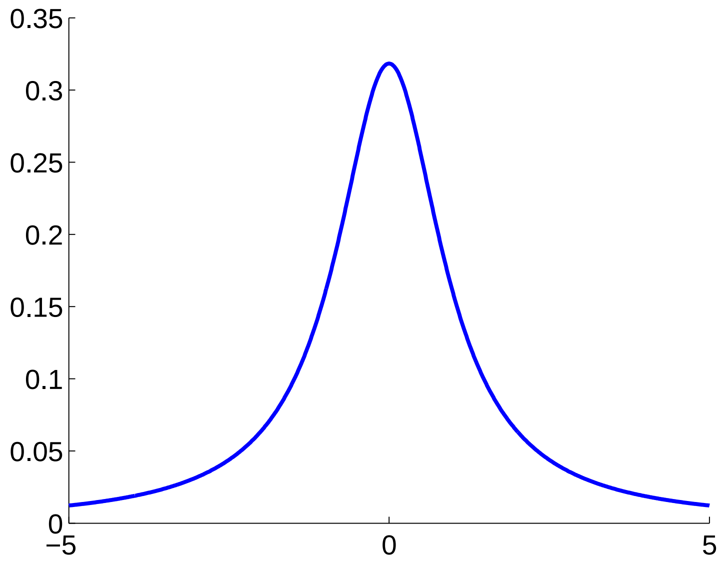


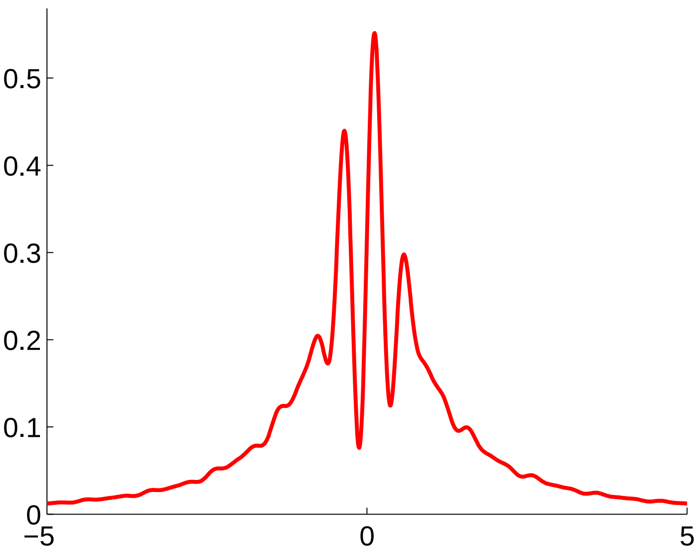
- $\theta_N(\omega) = \frac{A}{2j}\left[\odot_1^N \mathbf{1}_{[-M/2,M/2]}(\omega)\right] \odot [\delta(\omega-\omega_0) - \delta(\omega+\omega_0)]$; $(\mathbb{F}^{-1}\theta_N)(x) = \left(\frac{AM}{2\pi}\right)^N sinc^N\left(\frac{Mx}{2\pi}\right)$ with $|\omega_0| \geq \left(\frac{N+2}{2}\right) M$.





- Choose $q(x) = \frac{1}{\pi(1+x^2)}$, the Cauchy distribution.
- Construct $p(x) = q(x) + (\mathbb{F}^{-1}\theta_N)(x)$.





- Samples $\mathbf{X}$ and $\mathbf{Y}$ of size $n = 1000$ are drawn from p and q respectively.
- $\widehat{MMD}[\mathcal{F}, n, n]$ is verified to be 0.

Testing for Independence

- Testing independence between random variables X and Y can be posed as a two-sample problem based on the following result.

Theorem ([Jacod and Protter, 2000])

The random variables X and Y are independent if and only if $\mathbb{E}_{\mathbf{P}_{xy}}[f(x)g(y)] = \mathbb{E}_{\mathbf{P}_x \otimes \mathbf{P}_y}[f(x)g(y)]$ for each pair (f, g) of bounded continuous functions.

- Statistic for testing independence: $MMD[\mathcal{F} \otimes \mathcal{G}, \mathbf{P}_{xy}, \mathbf{P}_x \otimes \mathbf{P}_y]$.
- All the results derived before hold for independence testing also.

Summary

- RKHS based two-sample test can fail when:
  - a periodic kernel is used to test discrete probability measures on $\mathbb{R}^d$.
  - a kernel with uncountable holes in its spectrum is used to test non-discrete probability measures with non-compact support on $\mathbb{R}^d$.

References

Dudley, R. M. (2002).
Real Analysis and Probability.
Cambridge University Press, Cambridge, UK.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. (2007).
A kernel method for the two sample problem.
In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 513–520, Cambridge, MA. MIT Press.

Jacod, J. and Protter, P. (2000).
Probability Essentials.
Springer-Verlag, New York, USA.

Mallat, S. G. (1998).
A Wavelet Tour of Signal Processing.
Academic Press, San Diego.

Steinwart, I. (2002).
On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 2:67–93.

Strichartz, R. S. (2003).
A Guide to Distribution Theory and Fourier Transforms.
World Scientific Publishing, Singapore.