Mean Elements in RKHS

Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS

Department of Statistical Science, Graduate University for Advanced Studies

June 27, 2008 / Statistical Learning Theory II

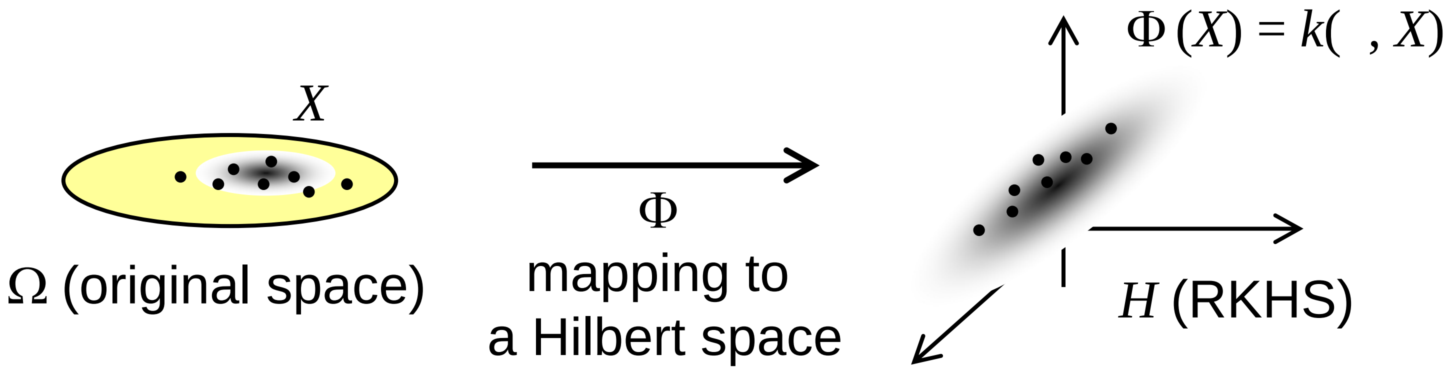
Outline

1. Introduction
2. Mean element in RKHS
3. Characteristic kernel
4. Summary

Introduction

- "Kernel methods" for statistical inference
  - We have seen that positive definite kernels are used for capturing 'nonlinearity' of original data through the higher-order moments.

    e.g. Support vector machine, kernel PCA, kernel CCA, etc.
  - Kernelization: mapping data into a RKHS and apply linear methods on the RKHS.



Do more basic descriptive statistics!

- Consider basic linear statistics (mean, variance, …) on RKHS, and their meaning on the original space.

- Basic statistics on Euclidean space / Basic statistics on RKHS

Basic statistics on Euclidean space	Basic statistics on RKHS
Mean	Mean element
Covariance	Cross-covariance operator
Conditional covariance	Conditional-covariance operator

Outline

1. Introduction
2. Mean element in RKHS
3. Characteristic kernel
4. Summary

Mean Element on RKHS I

$(\mathcal{X}, \mathcal{B})$: measurable sapce.

X: random variable taking value on $\mathcal{X}$.

k: measurable positive definite kernel on $\mathcal{X}$. H: RKHS defined by k.

$\Phi(X) = k(\cdot, X)$: random variable on RKHS.

- Assume $E\left[\sqrt{k(X,X)}\right] < \infty$. (satisfied by a bounded kernel)
- Define the mean element of X on H by $m_X \in H$ that satisfies

$$\langle m_X, f \rangle = E[f(X)] \qquad (\forall f \in H)$$

Existence and uniqueness:

$$\left| E[f(X)] \right| \le E\,|\langle f, k(\cdot, X)\rangle| \le \|f\|\, E\,\|k(\cdot, X)\| = E\left[\sqrt{k(X,X)}\right] \|f\|$$

$f \mapsto E[f(X)]$ is a bounded linear functional on H.

Use Riesz's lemma.

Mean Element on RKHS II

- Explicit form

$$m_X(u) = E[k(u, X)]$$

$$\because) \quad m_X(u) = \langle m_X, k(\cdot, u) \rangle = E[k(X, u)].$$

- Intuition on the role: the mean element contains the information of the higher-order moments.

  X: $\mathbf{R}$-valued random variable. k: pos.def. kernel on $\mathbf{R}$.

  Suppose pos. def. kernel k admits a power-series expansion on $\mathbf{R}$.

$$k(u, x) = c_0 + c_1(xu) + c_2(xu)^2 + \cdots \qquad (c_i > 0)$$

e.g.) $k(x, u) = \exp(xu)$

The mean element m_X works as a moment generating function:

$$m_X(u) = E[k(u, X)] = c_0 + c_1 E[X]u + c_2 E[X^2]u^2 + \cdots$$

$$\frac{1}{c_\ell} \frac{d^\ell}{du^\ell} m_X(u) \Big|_{u=0} = E[X^\ell]$$

Characteristic Kernel I

$\mathcal{P}$: family of all the probabilities on a measurable space $(\Omega, \mathcal{B})$.

H: RKHS on Ω with a bounded measurable kernel k.

m_P: mean element on H for a probability $P \in \mathcal{P}$

– Definition

The kernel k is called characteristic (w.r.t. $\mathcal{P}$) if the mapping

$$\mathcal{P} \to H, \qquad P \mapsto m_P$$

is one-to-one.

– The mean element for a characteristic kernel uniquely determines a probability.

$$m_P = m_Q \quad \Leftrightarrow \quad P = Q$$

i.e.

$$E_P[f(X)] = E_Q[f(X)] \ \ (\forall f \in \mathcal{H}) \quad \Leftrightarrow \quad P = Q.$$

Characteristic Kernel II

– Generalization of characteristic function

With Fourier kernel $k_F(x,y) = \exp\left(\sqrt{-1}\,x^T y\right)$

$$\mathrm{Ch.f.}_X(u) = E[k_F(X,u)].$$

- The characteristic function uniquely determines a Borel probability on $\mathbf{R}^m$.
- The mean element $m_X(u) = E[k(u,X)]$ w.r.t. a characteristic kernel uniquely determines a probability on $(\Omega, \mathcal{B})$.

  Note: Ω may not be Euclidean.

– The characteristic RKHS must be large enough!

Examples for $\mathbf{R}^m$ (proved later)

- Gaussian RBF kernel $\exp(-\frac{1}{2\sigma^2}\|x-y\|^2)$.
- Laplacian kernel $\exp(-\alpha\sum_{i=1}^m |x_i - y_i|)$.
- Polynomial kernels are not characteristic.

Determining a Probability I

- $\mathcal{P}$: family of all the probabilities on a measurable space $(\Omega, \mathcal{B})$.
- $\mathcal{F}$: a class of bounded measurable functions.

- When is the following map injective?

$$\mathcal{P} \to \mathcal{F}^*, \qquad P \mapsto \left(f \mapsto \int f dP\right)$$

i.e. $E_P[f(X)] = E_Q[f(X)] \ (\forall f \in \mathcal{F}) \quad \Rightarrow \quad P = Q.$

- $\mathcal{F} = \{I_E(x) \mid E \in \mathcal{B}\}$ (all index functions) satisfies this, of course.
- A characteristic RKHS is defined as such.
- For a metric space S, $\mathcal{F} = C_b(S)$ (Banach space of the bounded continuous functions) satisfies this.

Determining a Probability II

- Maximum mean discrepancy (MMD)

$$M(P,Q;\mathcal{F}) = \sup_{f\in\mathcal{F}} \left| E_{X\sim P}[f(X)] - E_{X\sim Q}[f(X)] \right|$$

– M is a distance on $\mathcal{P}$, if $\mathcal{F}$ satisfies the injective property.

– Let (H,k) be a RKHS. $\mathcal{F} = \{ f \in H \mid \| f \| \le 1 \}$.

$$M(P,Q;\mathcal{F}) = \sup_{\|f\|\le 1} \left| \left\langle f, m_P - m_Q \right\rangle \right| = \| m_P - m_Q \|$$

With a characteristic kernel k,

MMD = $\|m_P - m_Q\|$ is a distance over probabilities.

Empirical Estimation of Mean Element

Empirical mean element on RKHS

- An advantage of RKHS approach is its easy empirical estimation.
- $X^{(1)},\ldots,X^{(N)}$: i.i.d. sample → $\Phi(X_1),\ldots,\Phi(X_N)$: sample on RKHS

Empirical mean

$$\hat{m}_X^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\Phi(X_i) = \frac{1}{N}\sum_{i=1}^{N}k(\cdot, X_i)$$

The empirical mean element gives empirical average

$$\langle \hat{m}_X^{(N)}, f \rangle = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \quad \equiv \hat{E}_N[f(X)] \qquad (\forall f \in H)$$

Asymptotic Properties I

Theorem (strong $\sqrt{N}$-consistency)

Assume $E[k(X,X)] < \infty$.

$$\left\| \hat{m}_X^{(N)} - m_X \right\| = O_p\left(1/\sqrt{N}\right) \qquad (N \to \infty)$$

Proof.

$$E\|\widehat{m}^{(n)} - m_P\|^2 = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n E_{X_i} E_{X_j}[k(X_i, X_j)]$$
$$- \frac{2}{n}\sum_{i=1}^n E_{X_i} E_X[k(X_i, X)] + E_X E_{\tilde{X}}[k(X, \tilde{X})]$$
$$= \frac{1}{n^2}\sum_{i=1}^n \sum_{j\neq i} E[k(X_i, X_j)] + \frac{1}{n} E_X[k(X, X)] - E_X E_{\tilde{X}}[k(X, \tilde{X})]$$
$$= \frac{1}{n}\{E_X[k(X, X)] - E_X E_{\tilde{X}}[k(X, \tilde{X})]\}.$$

By Chebychev's inequality,

$$\Pr(\sqrt{n}\|\widehat{m}^{(n)} - m_X\| \geq \delta) \leq \frac{nE\|\widehat{m}^{(n)} - m_X\|^2}{\delta^2} = \frac{C}{\delta^2}. \qquad \square$$

Asymptotic Properties II

Corollary (Uniform law of large numbers)

Assume $E[k(X,X)] < \infty$.

$$\sup_{f\in H, \|f\|\le 1} \left| \frac{1}{N}\sum_{i=1}^{N} f(X_i) - E[f(X)] \right| = O_p(1/\sqrt{N}) \qquad (N \to \infty).$$

Proof.

$$LHS = \sup_{f\in H, \|f\|\le 1} \left| \langle \hat{m}_X^{(N)} - m_X, f\rangle \right| = \|\hat{m}_X^{(N)} - m_X\|.$$

$\square$

Asymptotic Properties III

Theorem (Convergence to Gaussian process)

Assume $E[k(X,X)] < \infty$.

$$\sqrt{N}(\widehat{m}^{(N)} - m_X) \quad \Rightarrow \quad G \quad \text{in law} \quad (N \to \infty),$$

where G is a centered Gaussian process on H with the covariance function

$$C(f,g) = E[f(X)g(X)] - E[f(X)]E[g(X)] = \mathrm{Cov}[f(X), g(X)].$$

Proof is omitted. See Berlinet & Thomas-Agnan, Theorem 108.

Application: Two-sample problem

- Homogeneity test

  Two i.i.d. samples are given;

  $$X^{(1)},\dots,X^{(N_X)} \quad \text{and} \quad Y^{(1)},\dots,Y^{(N_Y)}.$$

  Q: Are they sampled from the same distribution?

- Practically important.

  We often wish to distinguish two things:

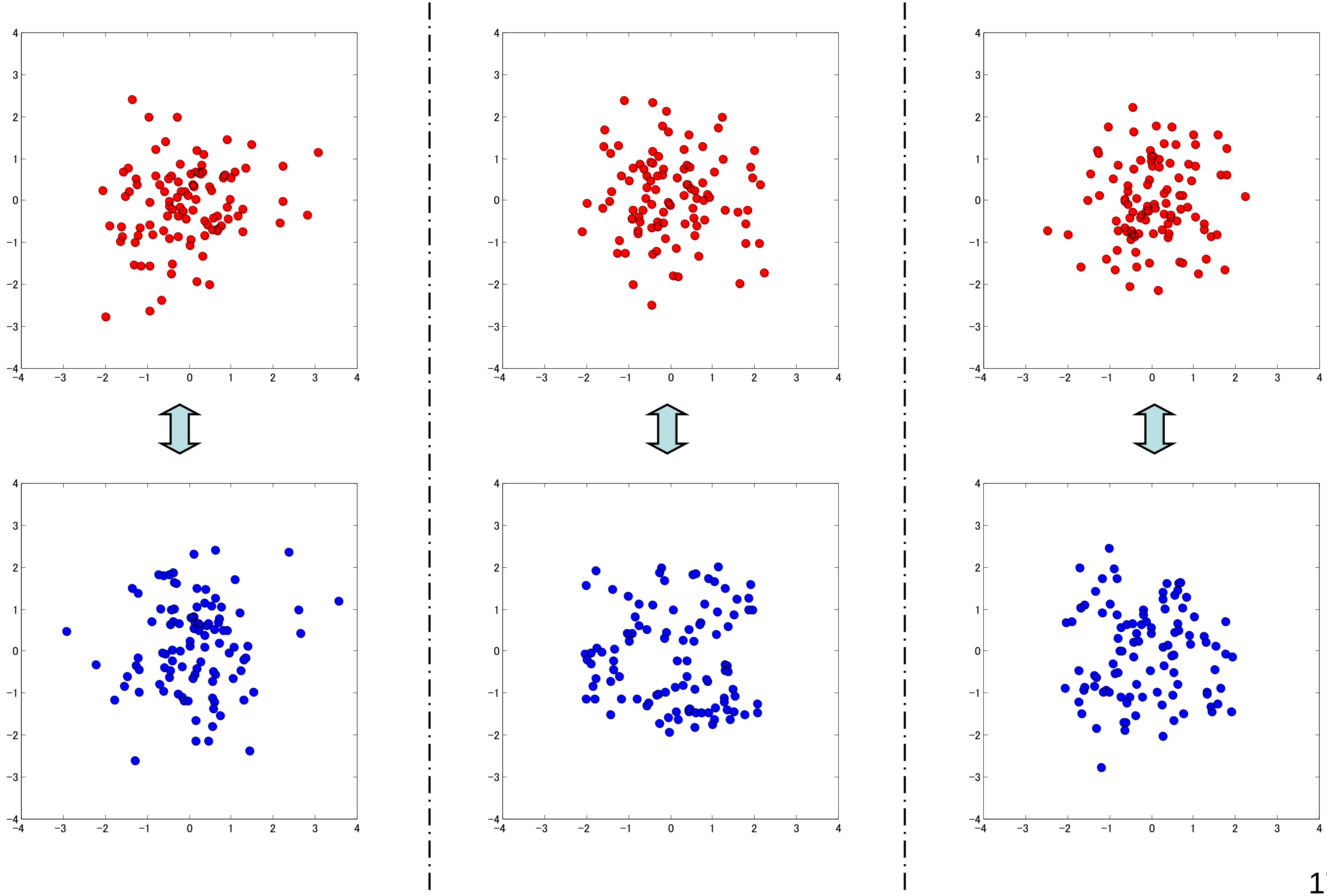
  - Are the experimental results of treatment and control significantly different?
  - Were the plays "*Henry VI*" and "*Henry II*" written by the same author?

- Kernel solution:

  Use the difference $m_X - m_Y$
  with a characteristic kernel such as Gaussian.

– Example: do they have the same distribution?

N = 100



Kernel Method for Two-sample Problem

Maximum Mean Discrepancy (Gretton etal 07, NIPS19)

– In population

$$MMD^2 = \|m_X - m_Y\|_H^2$$

– Empirically

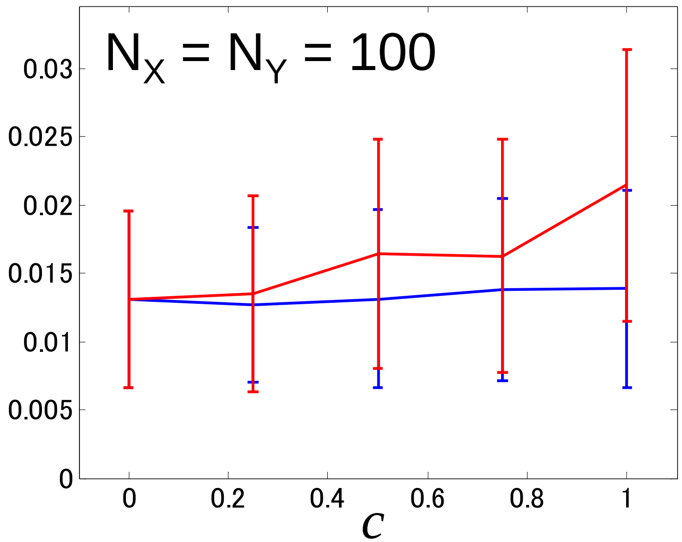
$$MMD_{emp}^2 = \|\hat{m}_X - \hat{m}_Y\|_H^2$$

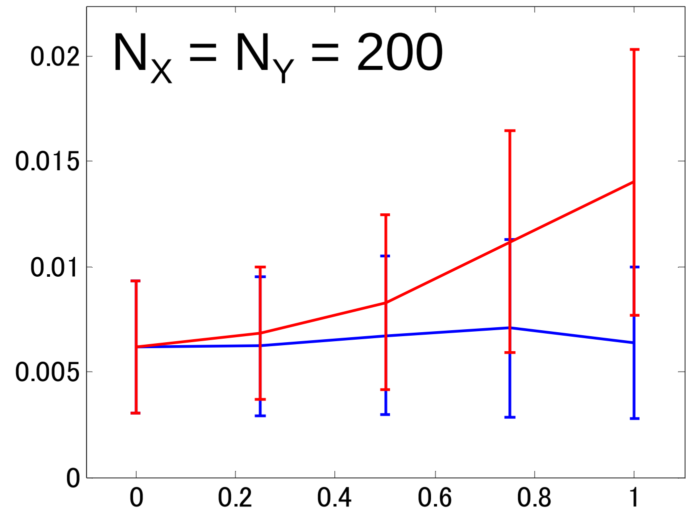
$$= \frac{1}{N_X^2}\sum_{i,j=1}^{N_X} k(X_i, X_j) - \frac{2}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{a=1}^{N_Y} k(X_i, Y_a) + \frac{1}{N_Y^2}\sum_{a,b=1}^{N_Y} k(Y_a, Y_b)$$

– With characteristic kernel, MMD = 0 if and only if $P_X = P_Y$.

– Asymptotic distribution of MMD_{emp}^2 is known, and used for two-sample homogeneity test (Gretton et al. 2007).

Experiment with MMD

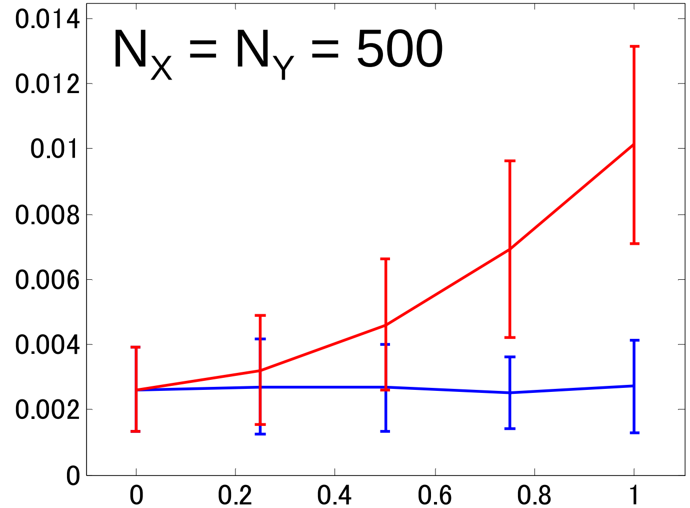




Means of MMD over 100 samples

N(0,1) vs
c Unif + (1-c) N(0,1)

N(0,1) vs N(0,1)



Outline

Conditions on Characteristic Kernels

Proposition

k: bounded measurable pos. def. kernel on a measurable space $(\Omega, \mathcal{B})$. H: associated RKHS. Then,

k is characteristic if and only if $H + \mathbf{R}$ is dense in $L^2(P)$ for any probability P on $(\Omega, \mathcal{B})$.

Proof.

$\Leftarrow$) Assume $m_P = m_Q$.

$|P - Q|$: the total variation of P - Q.

Since $H + \mathbf{R}$ is dense in $L^2(|P-Q|)$, for any $\varepsilon > 0$ and $A \in \mathcal{B}$ there exists $f \in H + \mathbf{R}$ and such that

$$\int |f - I_A| d(|P-Q|) < \varepsilon.$$

Thus, $|(E_P[f(X)] - P(A)) - (E_Q[f(X)] - Q(A))| < \varepsilon.$

From $m_P = m_Q$, $E_P[f(X)] = E_Q[f(X)]$, thus $|P(A) - Q(A)| < \varepsilon$.

This means $P = Q$.

$\Rightarrow$) Suppose H + $\mathbf{R}$ is not dense in $L^2(P)$.

There is $f \in L^2(P)\,(f \neq 0)$

$$\int f\varphi dP = 0, (\forall \varphi \in H), \quad \int f dP = 0.$$

Let $c = \frac{1}{\|f\|_{L^1(P)}}$.

Define probabilities Q_1 and Q_2 by

$$Q_1(E) = c\int_E |f| dP, \qquad Q_2(E) = c\int_E (|f| - f)dP.$$

$Q_1 \neq Q_2$ by $f \neq 0$.

But,

$$E_{Q_1}[k(\cdot, X)] - E_{Q_2}[k(\cdot, X)] = c\int f(x)k(\cdot, x)dP(x) = 0,$$

which means k is not characteristic.

$\square$

Shift-invariant Characteristic Kernels

– Continuous shift-invariant kernels on $\mathbf{R}^m$: $\phi(x\text{-}y)$

By Bochner's theorem, Fourier transform of ϕ is non-negative.

The characteristic kernels in this class are completely determined.

Theorem (Sriperumbudur et al. 2008)

Let $k(x,y) = \phi(x\text{-}y)$ be a $\mathbf{R}$-valued continuous shift-invariant positive definite kernel on $\mathbf{R}^m$ such that

$$\phi(x) = \int e^{\sqrt{-1}\omega^T x} d\Lambda(\omega).$$

Then, k is characteristic if and only if supp(Λ) = $\mathbf{R}^m$.

$$\mathrm{supp}(\mu) = \{x \in \mathbf{R}^m \mid \mu(U) \neq 0 \ \text{ for all open set } U \text{ s.t. } x \in U\}$$

Intuition:

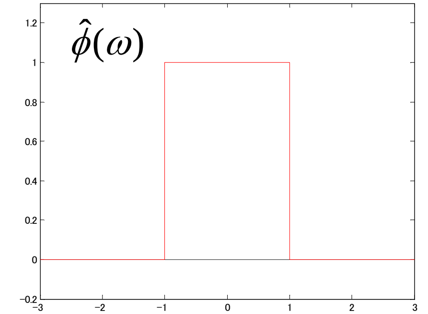
$$\int k(x-y)p(y)dx = \int k(x-y)q(y)dx \quad \Rightarrow \quad p = q$$

or

$$\hat{\phi}(\hat{p} - \hat{q}) = 0 \quad \Rightarrow \quad p = q$$

– Observation: if $\hat{\phi}(\omega) = 0$ on an interval of some frequency, then k must not be characteristic.

E.g. $\phi(x) = \dfrac{\sin(\alpha x)}{x}$ $\quad \hat{\phi}(\omega) = \sqrt{\frac{\pi}{2}}\, I_{[-\alpha\, \alpha]}(\omega)$

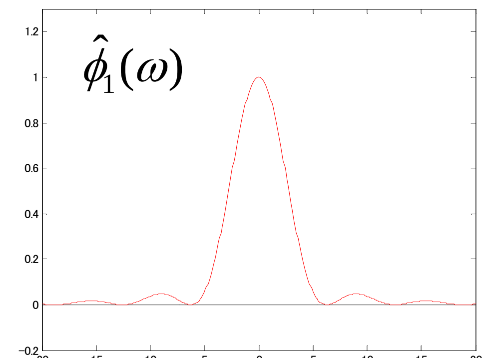


If $(p - q)^\wedge$ differ only out of $[-a, a]$,
p and q are not distinguishable.

– Conjecture: if $\hat{\phi}(\omega) > 0$ for all w, then $k(x, y) = f(x - y)$ is characteristic.

E.g. Gaussian kernel

$$\phi(x) = e^{-x^2/2\sigma^2} \qquad \hat{\phi}(\omega) = e^{-\sigma^2\omega^2/2}$$



– Is B_{2n+1}-spline kernel characteristic?

$$\phi_{2n+1}(x) = I_{[-\frac{1}{2}\ \frac{1}{2}]} * \cdots * I_{[-\frac{1}{2}\ \frac{1}{2}]}$$

$$\hat{\phi}_{2n+1}(\omega) = \left(\frac{2}{\pi}\right)^{n+1} \frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}}$$

– Examples

- Gaussian RBF kernels and Laplacican kernels are characteristic.

$$\phi(x) = e^{-x^2/2\sigma^2} \qquad \hat{\phi}(\omega) = e^{-\sigma^2\omega^2/2} \qquad \text{support} = \mathbf{R}$$

$$\phi(x) = e^{-\alpha|x|} \qquad \hat{\phi}(\omega) = \frac{2\alpha}{\pi(\alpha^2 + x^2)} \qquad \text{support} = \mathbf{R}$$

- B_{2n+1}-spline kernel is characteristic.

$$\hat{\phi}_{2n+1}(\omega) = \left(\frac{2}{\pi}\right)^{n+1} \frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}} \qquad \text{support} = \mathbf{R}$$

– Remark:
The Fourier analysis, Bochner’s theorem, and the theorem on shift-invariant characteristic kernels on $\mathbf{R}^m$ can be extended to locally compact Abelian groups.

Summary

- Mean element in RKHS
  - A random variable X can be transformed into a RKHS by

    $$\Phi(X) = k(\cdot, X)$$

    It contains the information of the higher-order moments of X.
  - The mean element is defined by $m_X = \mathrm{E}[\Phi(X)]$.
  - If the pos. def. kernel is characteristic, the mean element uniquely determines a probability.
  - The mean element with a characteristic kernel can be used for homogeneity tests.
  - The shift-invariant characteristic kernels on $\mathbf{R}^m$ (and locally compact Abelian groups) is completely determined.

Appendix: Hahn-Jordan decomposition I

Singed measure

$(\Omega, \mathcal{B})$: measurable space. $\mu : \mathcal{B} \to [-\infty, \infty]$.

μ is a signed measure if μ is countably additive, *i.e.*,

$$\mu(\emptyset) = 0,$$

$$\mu\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n),$$ $A_n (n = 1, 2, \ldots)$: disjoint.

Hahn-Jordan decomposition

Theorem.

For any signed measure μ on $(\Omega, \mathcal{B})$, there is $D \in \mathcal{B}$ such that

$$\mu^+(E) = \mu(E \cap D), \quad \mu^-(E) = -\mu(E \cap D^c) \qquad (\forall E \in \mathcal{B})$$

are non-negative measures.

$$\mu = \mu^+ - \mu^-.$$ (Jordan decomposition)

$$\mu^+(E) = \sup\{\mu(F) \mid F \subset E\}, \quad \mu^-(E) = -\inf\{\mu(F) \mid F \subset E\}.$$

Appendix: Hahn-Jordan decomposition II

Total variation

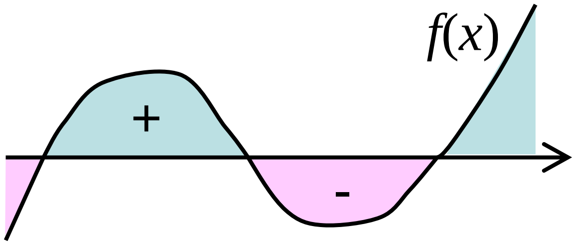
- For a signed measure μ

$$|\mu| \equiv \mu^+ + \mu^-. \qquad \text{(total variation)}$$

- If f is an integrable function on a measure space $(\Omega, \mathcal{B}, \nu)$,

$$\mu_f(E) = \int_E f(x) d\nu(x)$$

is a signed measure.



$$\mu_f^+(E) = \int_E f^+(x) d\nu(x),$$
$$\mu_f^-(E) = \int_E f^-(x) d\nu(x).$$

where

$$f^+(x) = \max\{f(x), 0\}, \qquad f^-(x) = \min\{f(x), 0\}.$$

The total variation is

$$|\mu_f|(E) = \int_E |f(x)| d\nu(x).$$

Appendix: Review of Fourier analysis

– Fourier transform of $f \in L^1(\mathbf{R}^\ell)$

$$\hat{f}(\omega) = \int f(x) e^{-\sqrt{-1}\omega^T x} dm_x \qquad dm_x = \frac{1}{(2\pi)^{\ell/2}} dx$$

– Fourier inverse transform

$$\check{F}(x) = \int F(\omega) e^{\sqrt{-1} x^T \omega} dm_\omega$$

– Fourier transform of a bounded $\mathbf{C}$-valued Borel measure μ

$$\hat{f}(\omega) = \int e^{-\sqrt{-1}\omega^T x} d\mu(x)$$

– Convolution

$$f * g = \int f(x-y) g(y) dy = \int g(x-y) f(y) dy$$

$$\mu * g = \int f(x-y) d\mu(y)$$

– Fourier transform of convolution:

$$(\mu * g)^\wedge = \hat{\mu}\, \hat{g}$$

References

Berlinet, A. and C. Thosma-Agnan. *Reproducing Kernel Hilbert Spaces in Probabiity and Statistics*. Kluwer Academic Press (2004).

Gretton, A., K. M. Borgwardt, M. Rasch, B. Schölkopf and A. Smola. A Kernel Method for the Two-Sample-Problem. *Advances in Neural Information Processing Systems 19*, 513-520, MIT Press. (2007)

Fukumizu, K., A. Gretton, X. Sun., and B. Schölkopf. Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 19*:489-496 (2008)

Sriperumbudur, B., A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert Space Embeddings of Probability Measures. *Proc. 21st Annual Conference on Learning Theory (COLT 2008).* (2008).