Kernel Method: Data Analysis with Positive Definite Kernels

7. Mean on RKHS and characteristic class

Kenji Fukumizu

The Institute of Statistical Mathematics
Graduate University for Advanced Studies /
Tokyo Institute of Technology

Nov. 17-26, 2010
Intensive Course at Tokyo Institute of Technology

Outline

1. Introduction

2. Mean on RKHS

3. Characteristic kernel

1. Introduction

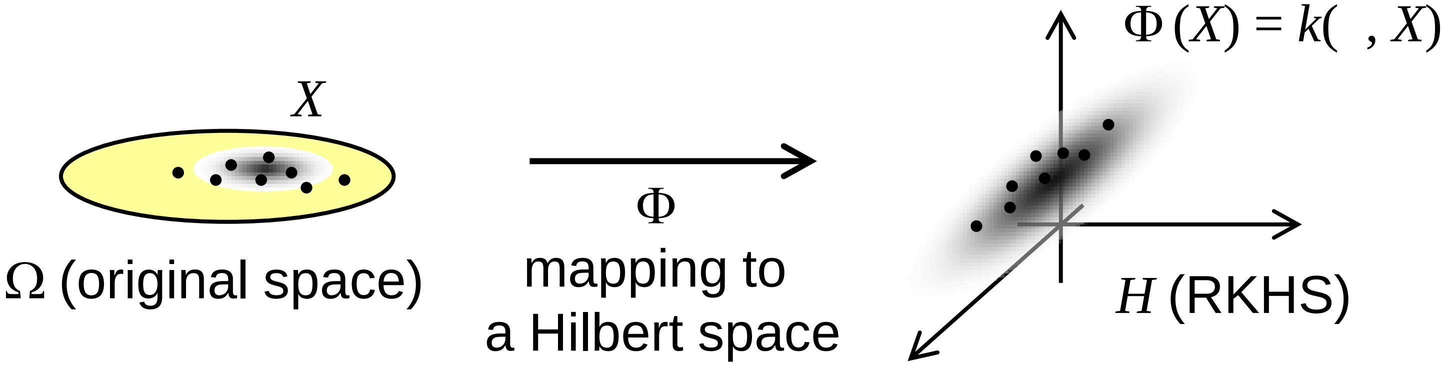
2. Mean on RKHS

3. Characteristic kernel

Introduction

- Kernel methods for statistical inference
  - We have seen that positive definite kernels are used for capturing ‘nonlinearity’ or ‘high-order moments’ of original data.

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



- Consider more basic statistics!
  - Consider basic 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
Covariance	Cross-covariance operator
Conditional covariance	Conditional-covariance operator

Outline

1. Introduction
2. Mean on RKHS
3. Characteristic kernel

Mean on RKHS I

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

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)

– We want to define the mean $E[\Phi(X)]$ of $\Phi(X)$ on H.

It can be defined as the integral of a Hilbert-valued function.

Mean on RKHS II

– Alternative definition:

Define the mean of X on H by $m_X \in H$ that satisfies

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

– Intuition:

Sample mean $\hat{m}_X(u) = \frac{1}{N}\sum_{i=1}^{N} \Phi(X_i)$

$$\langle \hat{m}_X, f \rangle = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \quad \Rightarrow \quad \langle m_X, f \rangle = E[f(X)]$$

– Explicit form:

$$m_X(u) = E[k(u,X)] = \int k(u,x)dP(x)$$

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

We call $m_X(u)$ kernel mean.

Mean on RKHS III

– Fact:

$$\langle E[k(\cdot,X)], f\rangle = E[\langle k(\cdot,X), f\rangle]$$

(exchangeability)

– The kernel mean does exists uniquely.

Existence and uniqueness:

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

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

Use Riesz's lemma.

Mean on RKHS IV

– Intuition: the mean contains the information of the high-order moments.

X: **R**-valued random variable. k: pos.def. kernel on **R**.

Suppose pos. def. kernel k admits a power-series expansion on **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 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 on H for a probability $P \in \mathcal{P}$

Def. 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 kernel mean by a characteristic kernel uniquely determines a probability.

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

i.e.

$$E_{X\sim P}[k(u,X)] = E_{X\sim Q}[k(u,X)] \quad \Leftrightarrow \quad P = Q$$

Characteristic Kernel II

– Generalization of characteristic function

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

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

- The characteristic function uniquely determines a Borel probability on $\mathbf{R}^m$.
- The kernel mean $m_X(u) = E[k(u,X)]$ by a characteristic kernel uniquely determines a probability on $(\Omega, \mathcal{B})$.
  Note: Ω may not be Euclidean.

Characteristic Kernel III

– The characteristic RKHS must be large enough!

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

- Gaussian RBF kernel

$$k_G(x, y) = \exp\left(-\frac{1}{2\sigma^2} \| x - y \|^2 \right)$$

- Laplacian kernel

$$k_L(x, y) = \exp\left(-\alpha \sum_{i=1}^{m} | x_i - y_i | \right)$$

- Polynomial kernels are not characteristic.
  - The RKHS for $(x^T y + c)^d$ is the space of polynomials of degree not greater than d.
  - The moments larger than d cannot be considered.

Empirical Estimation of Kernel Mean

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

Empirical kernel 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 kernel mean gives empirical average

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

Asymptotic Properties I

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

Assume $E[k(X,X)] < \infty$. For i.i.d. sample $X_1, \ldots, X_N$,

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

Proof.

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

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 $E[k(X,X)] < \infty$. For i.i.d. sample $X_1, \ldots, X_N$,

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

Proof.

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

□

Note: $\sup_{\|f\| \le 1} \left| \langle h, f \rangle \right| = \|h\|$

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

- Tow-sample homogeneity test

  Two i.i.d. samples are given;

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

  Q: Are they sampled from the same distribution?

- Practically important.

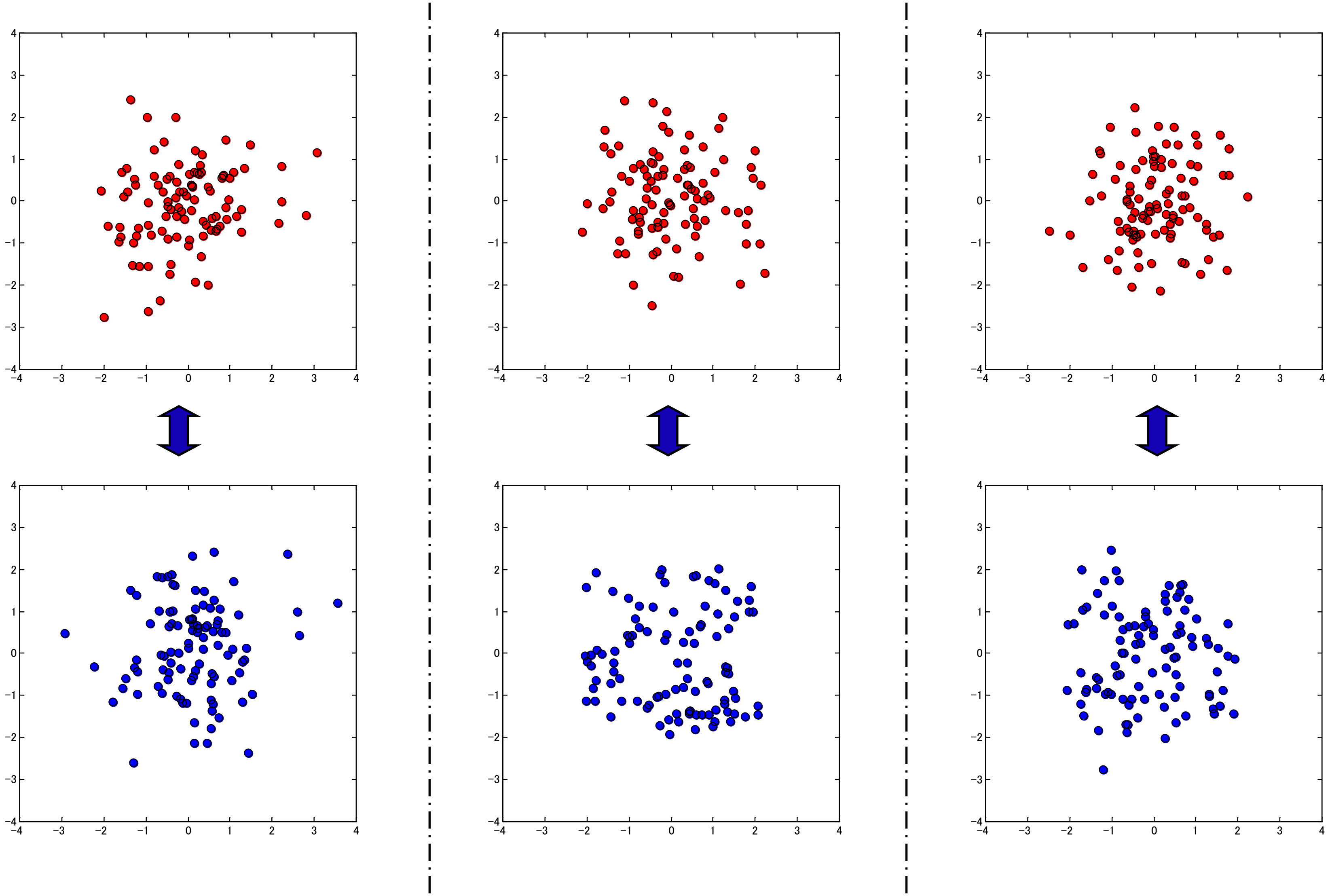
  We often wish to distinguish two things:

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

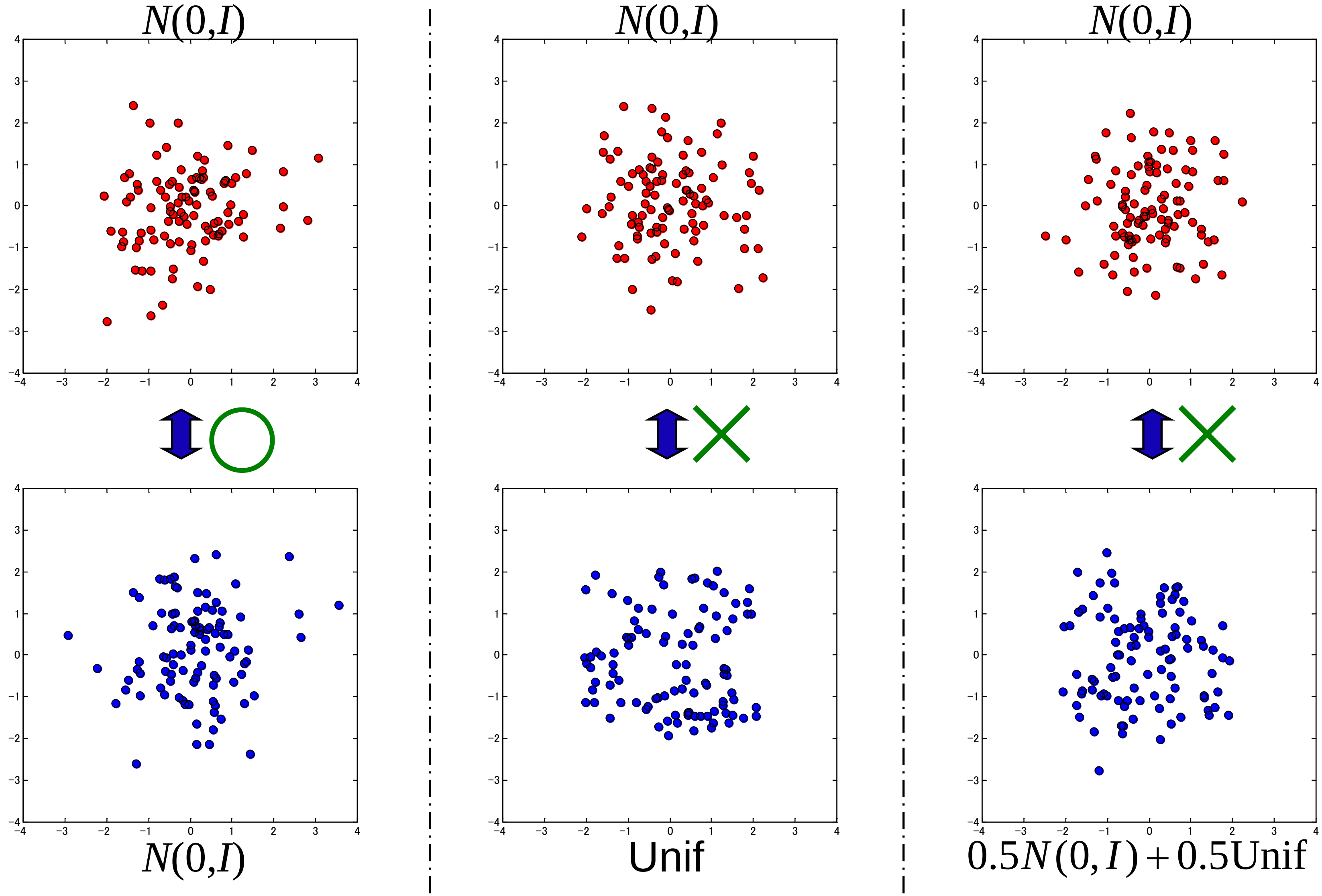
- Approach by kernel method: $m_X - m_Y$

  Use the difference of means with a characteristic kernel.

– Example: do they have the same distribution? N = 100



– Example: do they have the same distribution? N = 100



Kernel Method for Two-sample Problem

- Maximum Mean Discrepancy (Gretton et al 2007, 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. After debias, it is U-statistics.

Example

– Two sample test

P: $N(0,1/3)$ $\qquad$ Q_a: $a\phi(x;0,1/3)+(1-a)\frac{1}{2}I_{[-1,2]}(x).$

Null hypothesis H$_0$: $\mathrm{P}=Q_a$

Alternative H$_1$: $\mathrm{P}\neq Q_a$

– Results

- Comparison with Kolmogorov-Smirnov test
- Significance level = 5%. The asymptotic distribution is used.

	MMD					Kolmogorov-Smirnov				
N | a	1	0.75	0.5	0.25	0	1	0.75	0.5	0.25	0
200	0.966	0.898	0.788	0.964	0.882	0.962	0.910	0.730	0.956	0.940
500	0.990	0.868	0.544	0.118	0.038	0.990	0.752	0.382	0.112	0.124
1000	0.986	0.976	0.704	0.088	0	0.954	0.950	0.796	0.316	0.002

Percentage of accepting homogeneity in 500 simulations

1. Introduction

2. Mean element in RKHS

3. Characteristic kernel

Conditions on Characteristic Kernels I

Theorem (FBJ08+)

k: bounded measurable positive definite 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. See Appendix 1.

– The characteristic kernel must be large enough.

Def. A positive definite kernel on a compact space D is called universal if its RKHS is dense in $C(D)$.*

Proposition. A universal kernel is characteristic.

* $C(D)$ is the Banach space of the continuous function on D with sup norm.

Shift-invariant Characteristic Kernels II

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

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

The characteristic kernels in this class are completely determined.

– Intuition:

- For a shift-invariant kernel, the kernel mean is convolution:

$$m_P(u) = E_P[k(u, X)] = \int \phi(u - x)dP(x) = (\phi * p)(u)$$

- The characteristic property is equivalent to

$$\phi * p = \phi * q \quad \Rightarrow \quad p = q.$$

or by Fourier transform,

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

- It is expected that if $\hat{\phi}(\omega) > 0$ at any ω, then the above condition holds.

Shift-invariant Characteristic Kernels II

Theorem (Sriperumbudur et al. 2008)
Let $k(x,y) = \phi(x-y)$ be a **R**-valued continuous shift-invariant positive definite kernel on $\mathbf{R}^m$ such that

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

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

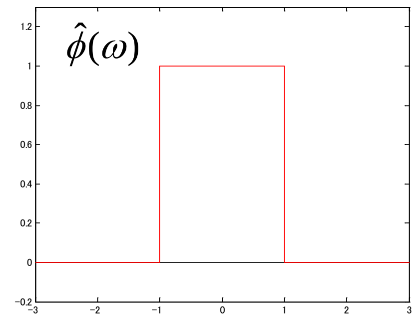
$$\mathrm{supp}(\mu) = \{x \in \mathbf{R}^m \mid \mu(U) \neq 0 \text{ for all open set } U \text{ s.t. } x \in U\}$$

Example

Gaussian	$\phi(x) = e^{-x^2/2\sigma^2}$	$\hat{\phi}(\omega) = e^{-\sigma^2\omega^2/2}$
Laplacian	$\phi(x) = e^{-\alpha\lvert x\rvert}$	$\hat{\phi}(\omega) = \dfrac{2\alpha}{\pi(\alpha^2 + x^2)}$
Cauchy	$\phi(x) = \dfrac{2\alpha}{\pi(\alpha^2 + x^2)}$	$\hat{\phi}(\omega) = e^{-\alpha\lvert\omega\rvert}$

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

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

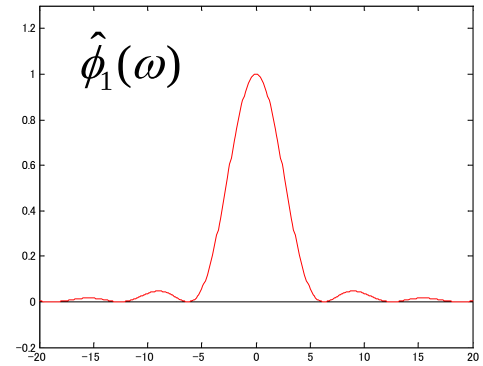


If $(p - q)^\wedge$ differ only out of [-a, a], p and q are not distinguishable.

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

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

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



– Bochner's theorem and the previous theorem can be extended to locally compact Abelian group.

Summary

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

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

    Its mean $m_X = E[\Phi(X)]$ contains the information of the higher-order moments of X.

  - If the positive definite kernel is characteristic, the kernel mean element uniquely determines a probability.

  - The kernel mean by characteristic kernel can be applied for two sample tests.

  - The shift-invariant characteristic kernels on $\mathbf{R}^m$ (and locally compact Abelian groups) is completely determined.

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 21*:489-496 (2008)

Fukumizu, K., B. Sriperumbudur, A. Gretton, B. Schölkopf. Characteristic Kernels on Groups and Semigroups. *Advances in Neural Information Processing Systems 22*: 2009 to appear.

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).

Appendix 1: proof on the characteristic kernel

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 $f \in L^2(P)\ (f \neq 0)$ such that

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

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

Define probabilities Q_1 and Q_2 by

$$Q_1(E) = c\int_E \left(|f(x)| - f(x)\right)dP(x), \quad Q_2(E) = c\int_E |f(x)|dP(x).$$

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

But,

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

which means k is not characteristic. □

Appendix 2: 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

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

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

– Fourier transform of convolution:

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

– Re: convolution $(f * g)^\wedge = \hat{f}\,\hat{g}$

Proof.

$$(f * g)^\wedge(\omega) = \int e^{-\sqrt{-1}x^T\omega} \int f(x-y)g(y)dm_y dm_x$$

$$= \int e^{-\sqrt{-1}(x-y)^T\omega} e^{-\sqrt{-1}y^T\omega} \int f(x-y)g(y)dm_y dm_x$$

$$= \int e^{-\sqrt{-1}z^T\omega} e^{-\sqrt{-1}y^T\omega} \int f(z)g(y)dm_y dm_z \qquad [z = x-y]$$

$$= \int e^{-\sqrt{-1}z^T\omega} f(z)dm_z \int e^{-\sqrt{-1}y^T\omega} g(y)dm_y$$

$$= \hat{f}(\omega)\hat{g}(\omega).$$