Independence, Conditional Independence, and Characteristic Kernels

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS

Graduate University for Advanced Studies

Based on joint work with Bharath Sriperumbudur, Arthur Gretton, Gert Lanckriet , and Bernhard Schölkopf

GIF Workshop @ Tübingen

May 15-16, 2008

Outline

1. Introduction
2. Characteristic kernels for determining probabilities
3. Shift-invariant characteristic kernels on locally compact Abelian groups
4. Summary

Introduction

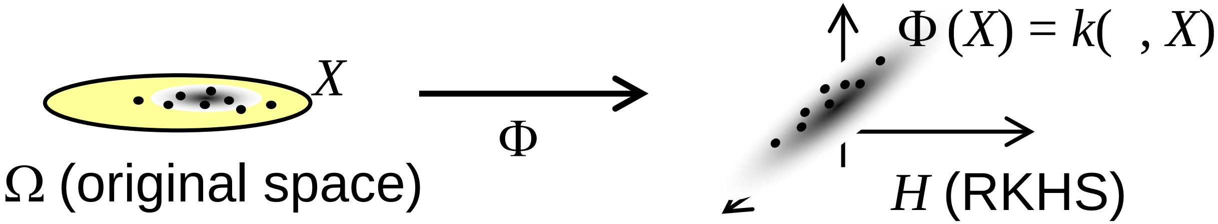
■ "Kernel methods" for statistical inference

- Kernelization: mapping *data* into a functional space (RKHS) and apply linear methods on RKHS.

- Transform the *random variable* X to $\Phi(X) = k(\cdot, X)$.

  Linear statistics on RKHS (variance, conditional covariance) can characterize independence and conditional independence through higher-order moments.

- With which kernels is this possible?



Outline

1. Introduction

2. Characteristic kernels for determining probabilities

3. Shift-invariant characteristic kernels on locally compact Abelian groups

4. Summary

Mean Element on RKHS

Mean element on RKHS

X: random variable taking value on Ω.

k: positive definite kernel on Ω. H: RKHS associated with k.

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

– There uniquely exists the mean element $m_X \in H$ of X on H s.t.

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

(by Riesz's lemma)

– Fact: $m_X(u) = E[k(u, X)]$

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

– m_X contains the information on the moments $E[f(X)]$ for all f.

If H is large enough, m_X may have sufficient information to determine the law of X

Determining Class

- Means determine a probability

  Proposition

  $(\Omega, \mathcal{B})$: measurable space. P, Q: probabilities on $(\Omega, \mathcal{B})$.

  If
  $$E_{X\sim P}[f(X)] = E_{X\sim Q}[f(X)]$$
  for every measurable function f, then, $P = Q$.

  Proposition (e.g. [Dudley 9.3.1])

  P, Q: Borel probabilities on a metric space.

  If
  $$E_{X\sim P}[f(X)] = E_{X\sim Q}[f(X)]$$
  for every continuous and bounded function f, then, $P = Q$.

  – The function class $C_b(\Omega)$ is a determining class of probabilities on a metric space.

Characteristic Kernels

- When does a RKHS work as a determining class?

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

  H: RKHS on Ω with measurable kernel k.

  m_P: mean element on H for a probability $P \in \mathcal{P}$ $\quad i.e.\ m_P(u) = E_P[k(X,u)]$

  – Definition: the kernel k is called characteristic 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.

  $$m_P(u) = m_Q(u) \quad (\forall u \in \Omega) \quad \Leftrightarrow \quad P = Q$$

  – Analogous to the characteristic function of a random vector

  $$\mathrm{Ch.f.}_X(u) = E[\exp^{\sqrt{-1}X^T u}].$$

Advantages of pos. def. kernel approach

– Empirical estimation is easy!

$X^{(1)},\ldots,X^{(N)}$: sample → $\Phi(X_1),\ldots,\Phi(X_N)$: sample on RKHS

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

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

– Application: 2-sample homogeneity test by MMD (Gretton et al. 2007)

$$MMD_{emp}^2 = \left\| \hat{m}_X - \hat{m}_Y \right\|_H^2$$
$$= \frac{1}{N_X^2}\sum_{i,j=1}^{N_X} k(X_i, X_j) - \frac{2}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{a=1}^{N_Y} k(X_i, Y_a) + \frac{1}{N_Y^2}\sum_{a,b=1}^{N_Y} k(Y_a, Y_b)$$

Statistical properties can be also derived.

Characterization of Independence

– Definition: cross-covariance operator

X , Y : general random variables on $\mathcal{X}$ and $\mathcal{Y}$, resp.

Prepare RKHS ($H_{\mathcal{X}}$, $k_{\mathcal{X}}$) and ($H_{\mathcal{Y}}$, $k_{\mathcal{Y}}$) defined on $\mathcal{X}$ and $\mathcal{Y}$, resp.

Define an operator $\Sigma_{YX} : H_X \to H_Y$

$$\langle g, \Sigma_{YX} f \rangle = E[g(Y) f(X)] - E[g(Y)]E[f(X)] \quad (= \mathrm{Cov}[f(X), g(Y)])$$

for all $f \in H_{\mathcal{X}}, g \in H_{\mathcal{Y}}$

– Independence and Cross-covariance operator

<u>Theorem</u>

If the product kernel $k_{\mathcal{X}} k_{\mathcal{Y}}$ is characteristic, then

X and Y are independent $\Leftrightarrow$ $\Sigma_{XY} = O$

- c.f. for Gaussian variables,

$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad V_{XY} = O$ *i.e.* uncorrelated

Characterization of Conditional Independence

X, Y, Z : random variables on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ (resp.).

$(H_{\mathcal{X}}, k_{\mathcal{X}}), (H_{\mathcal{Y}}, k_{\mathcal{Y}}), (H_{\mathcal{Z}}, k_{\mathcal{Z}})$: RKHS defined on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ (resp.).

– Conditional cross-covariance operator

$$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \qquad H_X \to H_Y$$

<u>Theorem (FBJ04, FBJ06, Sun et al 07)</u>

Define the augmented variable $\tilde{X} = (X, Z)$ and define a kernel on $\mathcal{X} \times \mathcal{Z}$ by

$$k_{\tilde{X}} = k_X k_Z$$

Assume $k_{\tilde{X}} k_Y$ and $k_{\mathcal{Z}}$ are characteristic, then,

$$\Sigma_{Y\tilde{X}|Z} = O \qquad \Leftrightarrow \qquad X \perp\!\!\!\perp Y \mid Z$$

c.f. for Gaussian variables,

$$V_{YY} - V_{YZ}V_{ZZ}^{-1}V_{ZX} = O \qquad \Leftrightarrow \qquad X \perp\!\!\!\perp Y \mid Z$$

Outline

1. Introduction
2. Characteristic kernels for determining probabilities
3. Shift-invariant characteristic kernels on locally compact Abelian groups
4. Summary

When is a kernel characteristic?

- Shift-invariant kernels on $\mathbf{R}^m$

<u>Bochner's theorem</u>

$\phi(x)$: bounded continuous function on $\mathbf{R}^m$.

A shift-invariant kernel $k(x, y) = \phi(x - y)$ is positive definite if and only if there is a non-negative finite Borel measure Λ such that

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

– If Λ is given by $\lambda(\omega)d\omega \;\; (\lambda(\omega) \geq 0)$

$$\lambda(\omega) = \hat{\phi}(\omega)$$ (Fourier transform of ϕ).

– Shift-invariant characteristic kernel on $\mathbf{R}^m$

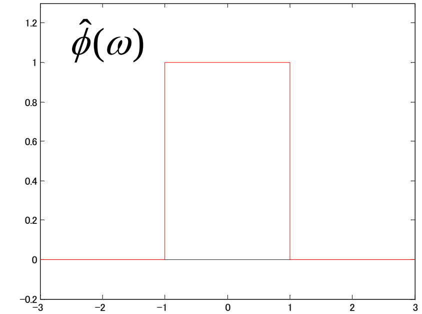
$$\int k(x-y)p(y)dx = \int k(x-y)q(y)dx \quad \Rightarrow \quad p = q$$

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

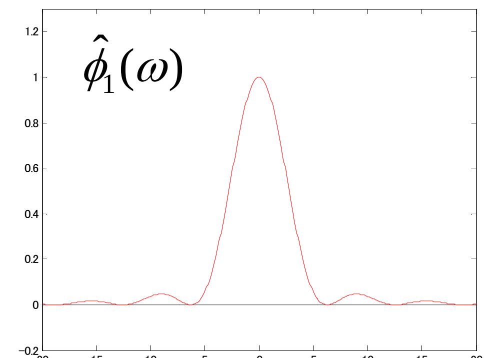


If $(p - q)^\wedge$ differ only out of $[-\alpha, \alpha]$,
p and q are not distinguishable.

– Conjecture: if $\hat{\phi}(\omega) > 0$ for all ω, then $k(x, y) = \phi(x - y)$ 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?

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

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

Locally Compact Abelian Group

– A Locally compact Abelian group (LCA group)

is a locally compact topological space with commutative group structure ($x + y = y + x$) such the group operations $(x, y) \mapsto x + y$ and $x \mapsto -x$ are continuous.

– Examples

- $\mathbf{R}^n$ with usual addition.
- $\mathbf{S}^1$ (unit circle) with addition modulo 2π.
- Torus: $\mathbf{S}^1$ x …x $\mathbf{S}^1$

– Haar measure: shift-invariant measure.

There is a unique (up to scale) Radon measure*

$\mu = dx$ on G s.t.

$$\mu(E + x) = \mu(E) \qquad (\forall x \in G, \forall E : \text{Borel set})$$

* A Radon measure is a Borel measure s.t. (i) $\mu(K) < \infty$ for all compact set K,
(ii) $\mu(E) = \sup\{\mu(K) \mid K \subset E, K : \text{compact}\} = \inf\{\mu(K) \mid E \subset U, U : \text{open}\}$

Fourier Analysis on LCA Group

– Character of LCA group

$\rho : G \to \mathbf{C}$: character of a LCA group G

$\underset{def.}{\Leftrightarrow}$ $|\rho(x)| = 1, \quad \rho(x+y) = \rho(x)\rho(y) \qquad (\forall x, y \in G)$

– Dual group: G^* = all the continuous characters on G.

The group operation is given by $(\rho\tau)(x) := \rho(x)\tau(x)$.

Examples

– $(\mathbf{R}^n, +)$: $G^* = \{e^{\sqrt{-1}\omega^T x} \mid \omega \in \mathbf{R}^n\}$ (Fourier kernels)

– $(\mathbf{S}^1, +)$: $G^* = \{e^{\frac{\sqrt{-1}n}{2\pi}x} \mid n \in \mathbf{Z}\}$ (Fourier kernels)

Fact: G^* is also a LCA group if the weakest topology so that $\rho \mapsto \rho(x)$ is continuous for every $x \in G$ is introduced.

Fact: $G^{**} \cong G$. (Pontryagin duality)

– On LCA group, Fourier analysis is possible by using the continuous characters as Fourier kernel.

- Fourier transform of $f \in L^1(G, dx)$

$$\hat{f}(\rho) = \int_G f(x)\overline{\rho(x)}dx \qquad \text{(function on } G^*\text{)}$$

- Fourier transform of a measure $\mu \in M(G)$.[1]

$$\hat{\mu}(\rho) = \int_G \overline{\rho(x)}d\mu(x)$$

- 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:

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

- Fourier inversion is also possible. $\breve{F}(x) = \int_{G^*} \rho(x)F(\rho)d\rho \quad (x \in G).$

1 $M(G)$ denotes the set of all bounded complex-valued Radon measures.

Bochner's Theorem

- Shift-invariant kernel on LCA group

  G: LCA group

  Shift-invariant positive definite kernel: $k(x, y) = \phi(x - y)$

<u>Bochner's theorem</u>

$\phi(x)$: bounded continuous function on a LCA group G.

The kernel $k(x, y) = \phi(x - y)$ is positive definite if and only if there is a non-negative measure $\Lambda \in M(G^*)$ such that

$$\phi(x) = \int_{G^*} \rho(x) d\Lambda(\rho) \qquad (x \in G).$$

The non-negative measure $\Lambda \in M(G^*)$ is unique.

$$\begin{array}{ccc} G & \underset{(\rho, x)}{\longleftrightarrow} & G^* \\ \phi & & \Lambda \end{array}$$

Shift-invariant Characteristic Kernels

– Support of a measure μ

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

Theorem (Sriperumbudur et al, COLT2008, Fukumizu et al. 2008)

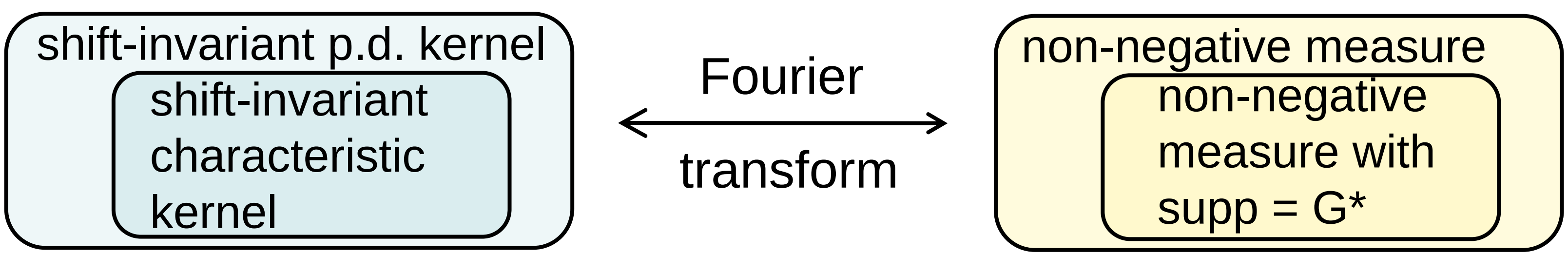
G: LCA group

$k(x, y) = \phi(x - y)$: shift-invariant positive definite kernel on G s.t.

$$\phi(x) = \int_{G^*} \rho(x) d\Lambda(\rho) \qquad (x \in G),$$

where Λ is a non-negative finite Borel measure on G^*.

k is characteristic if and only if $\mathrm{supp}(\Lambda) = G^*$.



– 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.

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

Summary

- Kernel methods for statistical inference
  - Transforming random variables into the feature space (RKHS).
  - Simple linear statistics on RKHS have rich information on the original variable.
  - To maintain all the information on the variables, use characteristic kernels.

- Shift-invariant characteristic kernels
  - Shift invariant characteristic kernels on a locally compact Abelian group can be determined completely by their Fourier transforms.