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.

\[ \Phi(X) = k(\ , X) \]

\[ \Omega \text{ (original space)} \]

\[ \Phi \text{ mapping to a Hilbert space} \]

\[ H \text{ (RKHS)} \]
• Consider more basic statistics!
  – Consider basic statistics (mean, variance, ...) on RKHS, and their meaning on the original space.
  
  – Basic statistics on Euclidean space
    Mean
    Covariance
    Conditional covariance
  
  – Basic statistics on RKHS
    Mean
    Cross-covariance operator
    Conditional-covariance operator
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 $\mathbb{E}[\sqrt{k(X,X)}] < \infty$. (satisfied by a bounded kernel)

– We want to define the mean $\mathbb{E}[\Phi(X)]$ 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

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

– Intuition:
  Sample mean
  $$\hat{m}_X(u) = \frac{1}{N} \sum_{i=1}^N \Phi(X_i) = \frac{1}{N} \sum_{i=1}^N k(\cdot, X_i)$$

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

– Explicit form:

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

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

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 exist uniquely.

Existence and uniqueness:

\[ |E[f(X)]| \leq E|\langle f, k(\cdot, X) \rangle| \leq \|f\| E\|k(\cdot, X)\| = E[\sqrt{k(X, X)}] \|f\|. \]

\[ f \mapsto E[f(X)] \] 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.

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

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

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

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

The mean $m_X$ works as a moment generating function:

$$m_X(u) = E[k(u, X)] = c_0 + c_1E[X]u + c_2E[X^2]u^2 + \cdots$$

$$\left. \frac{1}{c_\ell} \frac{d^\ell}{du^\ell} m_X(u) \right|_{u=0} = E[X^\ell]$$
Characteristic Kernel I

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

\( H \): RKHS on \( \Omega \) 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

\[
\mathcal{P} \rightarrow H, \quad P \mapsto m_P
\]

is one-to-one.

– The kernel mean by a characteristic kernel uniquely determines a probability.

\[
m_P = m_Q \iff P = Q
\]

i.e.

\[
E_{X \sim P}[k(u, X)] = E_{X \sim Q}[k(u, X)] \iff P = Q
\]
Characteristic Kernel II

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

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

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

- The characteristic RKHS must be large enough!

Examples for $\mathbb{R}^m$ (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 $(x^Ty + c)^d$ 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.
  - \(X^{(1)}, \ldots, X^{(N)}: \text{i.i.d. sample}\)
    \(\Rightarrow \Phi(X_1), \ldots, \Phi(X_N): \text{i.i.d. sample on RKHS}\)

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

**Theorem (strong $\sqrt N$ -consistency)**

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

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

**Proof.**

$$E\|\hat m_X^{(n)} - m_X\|^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} EX_i EX_j [k(X_i, X_j)]
- \frac{2}{n} \sum_{i=1}^{n} EX_i EX [k(X_i, X)] + EX EX [k(X, \tilde X)]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} E[k(X_i, X_j)] + \frac{1}{n} EX [k(X, X)] - EX EX [k(X, \tilde X)]$$

$$= \frac{1}{n} \{EX [k(X, X)] - EX EX [k(X, \tilde X)]\}.$$

By Chebychev’s inequality,

$$\Pr(\sqrt n\|\hat m^{(n)} - m_X\| \geq \delta) \leq \frac{nE\|\hat m^{(n)} - m_X\|^2}{\delta^2} = \frac{C}{\delta^2}. \quad \square$$
Corollary (Uniform law of large numbers)

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

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

Proof.

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

Note: \( \sup_{\|f\| \leq 1} \|\langle h, f \rangle\| = \|h\| \)
Theorem (Convergence to Gaussian process)

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

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

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

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

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
    \[ \text{MMD}^2 = \| m_X - m_Y \|_H^2 \]
  - Empirically
    \[ \text{MMD}_{\text{emp}}^2 = \| \hat{m}_X - \hat{m}_Y \|_H^2 \]
    \[= \frac{1}{N_X^2} \sum_{i,j=1}^{N_X} k(X_i, X_j) - \frac{2}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{a=1}^{N_Y} k(X_i, Y_a) + \frac{1}{N_Y^2} \sum_{a,b=1}^{N_Y} k(Y_a, Y_b) \]
  - With characteristic kernel, MMD = 0 if and only if \( P_X = P_Y \).
  - Asymptotic distribution of \( \text{MMD}_{\text{emp}}^2 \) is known. After debias, it is U-statistics.
Example

– Two sample test

\[ P: \ N(0,1/3) \]

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

Null hypothesis \( H_0: \ P = Q_a \)

Alternative \( H_1: \ P \neq Q_a \)

– Results

• Comparison with Kolmogorov-Smirnov test

• Significance level = 5%. The asymptotic distribution is used.

<table>
<thead>
<tr>
<th>( N )</th>
<th>( a )</th>
<th>( 1 )</th>
<th>0.75</th>
<th>0.5</th>
<th>0.25</th>
<th>0</th>
<th>( 1 )</th>
<th>0.75</th>
<th>0.5</th>
<th>0.25</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.966</td>
<td>0.898</td>
<td>0.788</td>
<td>0.964</td>
<td>0.882</td>
<td>0.962</td>
<td>0.910</td>
<td>0.730</td>
<td>0.956</td>
<td>0.940</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.990</td>
<td>0.868</td>
<td>0.544</td>
<td>0.118</td>
<td>0.038</td>
<td>0.990</td>
<td>0.752</td>
<td>0.382</td>
<td>0.112</td>
<td>0.124</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.986</td>
<td>0.976</td>
<td>0.704</td>
<td>0.088</td>
<td>0</td>
<td>0.954</td>
<td>0.950</td>
<td>0.796</td>
<td>0.316</td>
<td>0.002</td>
<td></td>
</tr>
</tbody>
</table>

Percentage of accepting homogeneity in 500 simulations
1. Introduction

2. Mean element in RKHS

3. Characteristic kernel
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 $H + \mathbb{R}$ 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

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

By Bochner’s theorem, Fourier transform of $\phi$ is non-negative. The characteristic kernels in this class are completely determined.

- Intuition:
  - For a shift-invariant kernel, the kernel mean is convolution:
    $m_P(u) = E_P[k(u, X)] = \int \phi(u - x)dP(x) = (\phi \ast p)(u)$
  
  - The characteristic property is equivalent to
    $\phi \ast p = \phi \ast q \implies p = q$.

  or by Fourier transform,
    $\hat{\phi} (\hat{p} - \hat{q}) = 0 \implies p = q$

  - It is expected that if $\hat{\phi}(\omega) > 0$ at any $\omega$, then the above condition holds.
**Theorem** (Sriperumbudur et al. 2008)

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

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

Then, \( k \) is characteristic if and only if \( \text{supp}(\Lambda) = \mathbb{R}^m \).

**Example on \( \mathbb{R} \)**

- **Gaussian**
  \[
  \phi(x) = e^{-x^2/2\sigma^2} \quad \hat{\phi}(\omega) = e^{-\sigma^2 \omega^2/2}
  \]

- **Laplacian**
  \[
  \phi(x) = e^{-\alpha|x|} \quad \hat{\phi}(\omega) = \frac{2\alpha}{\pi(\alpha^2 + \omega^2)}
  \]

- **Cauchy**
  \[
  \phi(x) = \frac{2\alpha}{\pi(\alpha^2 + x^2)} \quad \hat{\phi}(\omega) = e^{-\alpha|\omega|}
  \]
– if $\hat{\phi}(\omega) = 0$ on an interval of some frequency, then $k$ must not be characteristic.

E.g. $\phi(x) = \frac{\sin(\alpha x)}{x}$

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

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

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

$\phi_{2n+1}(x) = I_{[-\frac{1}{2}, \frac{1}{2}]} \ast \ldots \ast I_{[-\frac{1}{2}, \frac{1}{2}]}$

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

– 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 $\mathbb{R}^m$ (and locally compact Abelian groups) is completely determined.


Appendix 1: proof on the characteristic kernel

Proof.

\(\iff\) Assume \(m_P = m_Q\).

\(|P - Q|\): the total variation of \(P - Q\).

Since \(H + \mathbb{R}\) is dense in \(L^2(|P - Q|)\), for any \(\varepsilon > 0\) and \(A \in \mathcal{B}\) there exists \(f \in H + \mathbb{R}\) 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 \(P = Q\).
Suppose $H + R$ is not dense in $L^2(P)$.
There is $f \in L^2(P)$ ($f \neq 0$) such that
\[
\int f(x)\varphi(x)dP(x) = 0 \quad (\forall \varphi \in H), \quad \int f(x)dP(x) = 0.
\]
Let $c = 1/\|f\|_{L^1(P)}$.

Define probabilities $Q_1$ and $Q_2$ by
\[
Q_1(E) = c\int_E (|f(x)| - f(x))dP(x), \quad Q_2(E) = c\int_E |f(x)| dP(x).
\]

$Q_1 \neq Q_2$ 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. \qed
Appendix 2: Review of Fourier analysis

- Fourier transform of $f \in L^1(\mathbb{R}^\ell)$
  \[ \hat{f}(\omega) = \int f(x)e^{-\sqrt{-1} \omega^T x} \, dm_x \]
  \[ dm_x = \frac{1}{(2\pi)^{\ell/2}} \, dx \]

- Fourier inverse transform
  \[ \check{F}(x) = \int F(\omega)e^{\sqrt{-1} x^T \omega} \, dm_\omega \]

- Fourier transform of a bounded $\mathbb{C}$-valued Borel measure $\mu$
  \[ \hat{f}(\omega) = \int e^{-\sqrt{-1} \omega^T x} \, d\mu(x) \]

- Convolution
  \[ f \ast g = \int f(x-y)g(y) \, dm_y = \int g(x-y)f(y) \, dm_y \]
  \[ \mu \ast g = \int f(x-y) \, d\mu(y) \]

- Fourier transform of convolution:
  \[ (\mu \ast g)^\wedge = \hat{\mu} \hat{g} \]
Re: convolution \((f \ast g)^\hat{} = \hat{f} \hat{g}\)

Proof.

\[
(f \ast g)^\hat{}(\omega) = \int e^{-\langle x, \omega \rangle} \int f(x - y)g(y)dm_ydm_x
\]

\[
= \int e^{-\langle x-y, \omega \rangle} e^{-\langle y, \omega \rangle} \int f(x - y)g(y)dm_ydm_x
\]

\[
= \int e^{-\langle z, \omega \rangle} e^{-\langle y, \omega \rangle} \int f(z)g(y)dm_ydm_z \quad [z = x - y]
\]

\[
= \int e^{-\langle z, \omega \rangle} f(z)dm_z \int e^{-\langle y, \omega \rangle} g(y)dm_y
\]

\[
= \hat{f}(\omega)\hat{g}(\omega).
\]