Dependence Analysis with Reproducing Kernel Hilbert Spaces

Kenji Fukumizu
Institute of Statistical Mathematics
Graduate University for Advanced Studies

Based on collaborations with M. Jordan (UC Berkeley), F. Bach (Ecole Normale Supérieure), A. Gretton, X. Sun, and B. Schölkopf (Max-Planck Institute)

7th World Congress on Statistics and Probability
July 14-19, 2008. Singapore
Outline

- Introduction
- Independence and conditional independence with RKHS
- Kernel dimension reduction for regression
- Summary
RKHS for statistical inference

“RKHS methods” for statistical inference

- Reproducing kernel Hilbert space (RKHS) / positive definite kernel: capture “nonlinearity” or “higher-order moments” of data.
  e.g. Support vector machine.

- Recent studies:
  RKHS applied to independence and conditional independence.

\[ \Phi \rightarrow H_X \rightarrow \text{RKHS} \]

Apply linear methods on the transformed data
Positive definite kernel and RKHS

■ Positive definite kernel

Ω: set. \( k: \Omega \times \Omega \rightarrow \mathbb{R} \)

\( k \) is positive definite if \( k(x,y) = k(y,x) \) and for any \( n \in \mathbb{N}, \ x_1, \ldots, x_n \in \Omega \)
the matrix \( (k(x_i, x_j))_{i,j} \) (Gram matrix) is positive semidefinite.

- Example: Gaussian RBF kernel \( k(x,y) = \exp\left(-\|x-y\|^2/\sigma^2\right) \)

■ Reproducing kernel Hilbert space (RKHS)

\( k \): positive definite kernel on \( \Omega \).

\( \exists \mathcal{H}: \) Hilbert space consisting of functions on \( \Omega \) such that

1) \( k(\cdot, x) \in \mathcal{H} \) for all \( x \in \Omega \).

2) \( \text{Span}\{k(\cdot, x) | x \in \Omega\} \) is dense in \( \mathcal{H} \).

3) \( \langle k(\cdot, x), f \rangle_{\mathcal{H}} = f(x) \quad \forall f \in \mathcal{H}, x \in \Omega. \) (reproducing property)
How to use RKHS for data analysis?

Transform data into RKHS.

\[ \Phi : \Omega \rightarrow \mathcal{H}, \quad x \mapsto k(\cdot, x) \]

i.e. \[ \Phi(x) = k(\cdot, x) \]

Data: \( X_1, \ldots, X_N \) \( \rightarrow \) \( \Phi(X_1), \ldots, \Phi(X_N) \) : functional data

Illustration of dependence analysis with RKHS
Why RKHS? Easy empirical computation

The inner product of $\mathcal{H}$ is efficiently computable, while the dimensionality may be infinite.

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

$$f = \sum_{i=1}^{N} a_i \Phi(x_i), \quad g = \sum_{j=1}^{N} b_j \Phi(x_j) \quad \Rightarrow \quad \langle f, g \rangle = \sum_{i,j=1}^{N} a_i b_j k(x_i, x_j)$$

- The computational cost essentially depends on the sample size $N$.
- *c.f.* $L^2$ inner product / power expansion

$$(X, Y, Z, W) \mapsto (X, Y, Z, W, X^2, Y^2, Z^2, W^2, XY, XZ, XW, YZ, \ldots)$$

- Advantageous for high-dimensional data of moderate sample size.
- Can be applied for non-Euclidean data (strings, graphs, etc.).
Outline

- Introduction

- Independence and conditional independence with RKHS

- Kernel dimension reduction for regression

- Summary
Covariance on RKHS

\((X, Y)\): random vector taking values on \(\Omega_X \times \Omega_Y\).

\((\mathcal{H}_X, k_X), (\mathcal{H}_Y, k_Y)\): RKHS on \(\Omega_X\) and \(\Omega_Y\), resp.

Define random variables on the RKHS \(\mathcal{H}_X\) and \(\mathcal{H}_Y\) by

\[
\Phi_X(X) = k_X(\cdot, X), \quad \Phi_Y(Y) = k_Y(\cdot, Y).
\]

Def. Cross-covariance operator \(\Sigma_{YX}: \mathcal{H}_X \rightarrow \mathcal{H}_Y\)

\[
\Sigma_{YX} = E[\Phi_Y(Y) \otimes \Phi_X(X)] - E[\Phi_Y(Y)] \otimes E[\Phi_X(X)]
\]

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

for all \(f \in \mathcal{H}_X, g \in \mathcal{H}_Y\)

Characterization of independence

- Independence and cross-covariance operator
  If the RKHS’s are “rich enough” to express all the moments,

  $X \perp Y \iff \Sigma_{XY} = O \iff E[g(Y)f(X)] = E[g(Y)]E[f(X)]$
  for all $f \in \mathcal{H}_X, g \in \mathcal{H}_Y$

  $f$ and $g$ are test functions to compare the moments with respect to $P_{XY}$ and $P_XP_Y$.

  - Analog to Gaussian random vectors: $X \perp Y \iff V_{YX} = O$.
  
  - C.f. characteristic function
    $X \perp Y \iff E_{XY}[e^{\sqrt{-1}\omega^T X}e^{\sqrt{-1}\eta^T Y}] = E_X[e^{\sqrt{-1}\omega^T X}]E_Y[e^{\sqrt{-1}\eta^T Y}]$
    for all $\omega$ and $\eta$.

  - Applied to independence test (Gretton et al. 2008).
Characteristic kernels

A class for determining a probability

$X$: random variable taking values on $\Omega$.

$(\mathcal{H}, k)$: RKHS on $\Omega$ with a bounded measurable kernel $k$.

$\mathcal{H}$ (or $k$) is called characteristic if, for probabilities $P$ and $Q$ on $\Omega$,

$$E_{X \sim P}[f(X)] = E_{X \sim Q}[f(X)] \quad (\forall f \in \mathcal{H}) \text{ means } P = Q.$$  

($\mathcal{H}$ works as a class of test functions to determine a probability.)

- If $\mathcal{H}_X \otimes \mathcal{H}_Y$ given by the product kernel $k_X k_Y$ is characteristic,

$$X \perp Y \iff \Sigma_{XY} = O.$$  

($\Sigma_{XY} = O \iff E_{P_{XY}}[f(X)g(Y)] = E_{P_X P_Y}[f(X)g(Y)] \iff P_{XY} = P_X P_Y.$)

- An example on $\mathbb{R}^m$: Gaussian RBF kernel $\exp\left(-\|x - y\|^2 / \sigma^2\right)$
Estimation of cross-cov. operator

\((X_1, Y_1), \ldots, (X_N, Y_N)\) : i.i.d. sample on \(\Omega_X \times \Omega_Y\).

\[\hat{\Sigma}^{(N)}_{YX} = \frac{1}{N} \sum_{i=1}^{N} k_Y(\cdot, Y_i) \otimes k_X(\cdot, X_i) - \left( \frac{1}{N} \sum_{i=1}^{N} k_Y(\cdot, Y_i) \right) \otimes \left( \frac{1}{N} \sum_{i=1}^{N} k_X(\cdot, X_i) \right).\]

\((\text{rank} \leq N)\)

\[\langle g, \hat{\Sigma}^{(N)}_{YX} f \rangle = \frac{1}{N} \sum_{i=1}^{N} g(Y_i) f(X_i) - \left\{ \frac{1}{N} \sum_{i=1}^{N} g(Y_i) \right\} \left\{ \frac{1}{N} \sum_{i=1}^{N} f(X_i) \right\}.\]

\(\hat{\Sigma}^{(N)}_{YX}\) is represented by the Gram matrices.

<table>
<thead>
<tr>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>(|\hat{\Sigma}^{(N)}<em>{YX} - \Sigma</em>{YX}|_{HS} = O_p \left(1/\sqrt{N}\right)) (N \to \infty)</td>
</tr>
</tbody>
</table>

– A uniform law of large numbers follows:

\[\sup_{\|f\|_{H_X} \leq 1, \|g\|_{H_Y} \leq 1} \left| \text{Cov}_{emp}[f(X), g(Y)] - \text{Cov}[f(X), g(Y)] \right| \to 0 \quad \text{in pr.} \quad (N \to \infty).\]

– Weak convergence of \(\sqrt{N}(\hat{\Sigma}^{(N)}_{YX} - \Sigma_{YX})\) to a Gaussian process on \(\mathcal{H}_X \otimes \mathcal{H}_Y\) is also known.
RKHS and conditional independence

- **Conditional covariance operator**
  
  $X$ and $Y$: random variables. $\mathcal{H}_X, \mathcal{H}_Y$: RKHS with kernel $k_X, k_Y$, resp.

  Def. $\Sigma_{Y|X} \equiv \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$: **conditional covariance operator** on $\mathcal{H}_Y$

  (Analogous to conditional covariance matrix $V_{YY} - V_{YX}V_{XX}^{-1}V_{XY}$)

  - Relation to conditional variance:
    
    If $k_X$ is characteristic (e.g. Gaussian RBF kernel),

    $$\langle g, \Sigma_{Y|X}g \rangle = E[\text{Var}[g(Y) \mid X]] = \inf_{f \in \mathcal{H}_X} E[(g(Y) - E[g(Y)]) - (f(X) - E[f(X)])]^2$$

    ($\forall g \in \mathcal{H}_Y$)

  - Empirical estimator

    $$\hat{\Sigma}_{Y|X}^{(N)} = \hat{\Sigma}_{YY}^{(N)} - \hat{\Sigma}_{YX}^{(N)}\left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon_N I\right)^{-1}\hat{\Sigma}_{XY}^{(N)}$$

    $\varepsilon_N$: regularization coefficient

    Can be represented by Gram matrices.


Conditional independence

**Theorem (FBJ 2004, 2006)**

\(U, V,\) and \(Y\) are random variables on \(\Omega_U,\ \Omega_V,\) and \(\Omega_Y,\) resp.

\(H_U,\ H_V,\ H_Y:\) RKHS on \(\Omega_U,\ \Omega_V,\ \Omega_Y\) with kernel \(k_U,\ k_V,\ k_Y,\) resp.

\(X = (U, V).\) RKHS on \(\Omega_X = \Omega_U \times \Omega_V\) is defined by \(k_X = k_U k_V.\)

Assume \(H_X,\ H_U:\) characteristic. Then,

\[
\Sigma_{YY|U} \geq \Sigma_{YY|X} \geq: \text{the partial order of self-adjoint operators}
\]

If further \(H_Y\) is characteristic, then

\[
Y \perp X | U \iff \Sigma_{YY|U} = \Sigma_{YY|X}
\]

\[
\text{Tr}[\Sigma_{YY|U} - \Sigma_{YY|X}] \text{ works as a measure of conditional independence.}
\]

\(B \geq A\) means that \(B - A\) is positive semidefinite.
Outline

- Introduction
- Independence and conditional independence with RKHS
- Kernel dimension reduction for regression
- Summary
Dimension reduction for regression

- Regression: \( Y \) : response variable,
  \( X=(X_1,...,X_m) \): \( m \)-dim. explanatory variable

- Goal of dimension reduction for regression
  = Find an effective direction for regression (EDR space)

  \[
  p(Y \mid X) = \tilde{p}(Y \mid b_1^T X, ..., b_d^T X) \quad \left( = \tilde{p}(Y \mid B^T X) \right)
  \]
  \[
  B=(b_1, ..., b_d): \quad m \times d \quad \text{matrix} \quad d \text{ is fixed.}
  \]

  \[\iff\] \( X \perp Y \mid B^T X \)

- Existing methods:
  Sliced Inverse Regression (SIR, Li 1991),
  principal Hessian direction (pHd, Li 1992),
  SAVE (Cook&Weisberg 1991),  MAVE (Xia et al 2002),
  contour regression (Li et al 2005), among others.
Kernel Dimension Reduction
(Fukumizu, Bach, Jordan 2004, 2006)

Use characteristic kernels for $B^TX$ and $Y$.

$$\Sigma_{YY|B^TX} \geq \Sigma_{YY|X}$$

$$\Sigma_{YY|B^TX} = \Sigma_{YY|X} \iff X \perp Y \mid B^TX \quad \text{EDR space}$$

- KDR objective function

$$\min_{B: B^T B = I_d} \text{Tr} \left[ \Sigma_{YY|B^TX} \right]$$

- KDR contrast function with finite sample

$$\min_{B: B^T B = I_d} \text{Tr} \left[ G_Y \left( G_{B^TX} + N \varepsilon_N I_N \right)^{-1} \right]$$

where

$$G_{B^T X} = \left( I_N - \frac{1}{N} 1_N 1_N^T \right) K_{B^T X} \left( I_N - \frac{1}{N} 1_N 1_N^T \right): \text{centered Gram matrix}$$

$$K_{B^T X, ij} = k_d \left( B^T X_i, B^T X_j \right)$$
Wide applicability of KDR

- The most general approach to dimension reduction:
  - no model is used for \( p(Y|X) \) or \( p(X) \).
  - no strong assumptions on the distribution of \( X, Y \) and
dimensionality/type of \( Y \).
- Most conventional methods have some restrictions.

Computational issues

- Computational cost with matrices of sample size.
  \( \rightarrow \) Low-rank approximation, e.g. incomplete Cholesky
decomposition.
- Non-convex contrast function, possibly local minima.
  \( \rightarrow \) Gradient method with an annealing technique
    starting from a large \( \sigma \) in Gaussian RBF kernel.
Consistency of KDR

Theorem (FBJ2006)

Suppose $k_d$ is bounded and continuous, and

$$
\varepsilon_N \to 0, \ N^{1/2}\varepsilon_N \to \infty \ (N \to \infty).
$$

Let $S_0$ be the set of the optimal parameters;

$$
S_0 = \left\{ B \mid B^T B = I_d, \ \text{Tr}\left[ \Sigma_{Y|B^TX} \right] = \min_{B'} \text{Tr}\left[ \Sigma_{Y|B'^TX} \right] \right\}
$$

Estimator: \( \hat{B}^{(N)} = \min_{B : B^T B = I_d} \ \text{Tr}\left[ G_Y (G_{B^TX} + N\varepsilon_N I_N)^{-1} \right] \)

Then, under some conditions, for any open set $U \supset S_0$

$$
\Pr\left( \hat{B}^{(N)} \in U \right) \to 1 \quad (N \to \infty).
$$
Numerical results with KDR

- **Synthetic data (A)**

  \[ X : 4 \text{ dim. } \sim N(0, I_4) \]

  \[ Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + (1 + X_2)^2 + W. \quad W \sim N(0, \tau^2). \quad \tau = 0.1, 0.4, 0.8. \]

  Sample size \( N = 100 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>KDR Mean</th>
<th>KDR SD</th>
<th>SIR Mean</th>
<th>SIR SD</th>
<th>SAVE Mean</th>
<th>SAVE SD</th>
<th>pHd Mean</th>
<th>pHd SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.11</td>
<td>\pm 0.07</td>
<td>0.55</td>
<td>\pm 0.28</td>
<td>0.77</td>
<td>\pm 0.35</td>
<td>1.04</td>
<td>\pm 0.34</td>
</tr>
<tr>
<td>0.4</td>
<td>0.17</td>
<td>\pm 0.09</td>
<td>0.60</td>
<td>\pm 0.27</td>
<td>0.82</td>
<td>\pm 0.34</td>
<td>1.03</td>
<td>\pm 0.33</td>
</tr>
<tr>
<td>0.8</td>
<td>0.34</td>
<td>\pm 0.22</td>
<td>0.69</td>
<td>\pm 0.25</td>
<td>0.94</td>
<td>\pm 0.35</td>
<td>1.06</td>
<td>\pm 0.33</td>
</tr>
</tbody>
</table>

Frobenius norms of the projection matrices over 100 samples.
(Means and standard deviations)
**Synthetic data (B)**

\[ X : 10 \text{ dim. } \sim N(0, I_4) \]

\[ Y = \frac{1}{2} (X_1 - a)^2 W. \quad W \sim N(0,1). \quad a = 0, \ 0.5, \ 1. \]

Sample size \( N = 500 \)

<table>
<thead>
<tr>
<th></th>
<th>KDR</th>
<th>SIR</th>
<th>SAVE</th>
<th>pHd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.17</td>
<td>±0.05</td>
<td>1.83</td>
<td>±0.22</td>
</tr>
<tr>
<td>0.5</td>
<td>0.17</td>
<td>±0.04</td>
<td>0.58</td>
<td>±0.19</td>
</tr>
<tr>
<td>1.0</td>
<td>0.18</td>
<td>±0.05</td>
<td>0.30</td>
<td>±0.08</td>
</tr>
</tbody>
</table>
KDR on Real data

**Wine data**

- Data
  - 13 dim. 178 data
  - 3 classes
  - 2 dim. projection

\[
k(z_1, z_2) = \exp\left(-\frac{\|z_1 - z_2\|^2}{\sigma^2}\right)
\]

\[\sigma = 30\]
Swiss bank notes data

$X$: 6 dim. (measurements of each bank note)

$Y$: binary (genuine/counterfeit)

100 counterfeits and 100 genuine notes

$$k(z_1, z_2) = \exp(-\frac{||z_1 - z_2||^2}{a})$$

SAVE
Summary

- Positive definite kernels give a nice tool for dependence analysis
  - Covariance and conditional covariance operators on RKHS characterize independence and conditional independence.

- Kernel dimension reduction for regression (KDR)
  - The most general approach to dimension reduction.

- Future/ongoing studies
  - Choice of dimensionality for KDR.
  - Further asymptotic properties of the KDR estimator.

