Kernel Method: Data Analysis with Positive Definite Kernels

8. Dependence analysis with covariance on RKHS

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. Covariance operators on RKHS

2. Independence and dependence with kernels

3. Conditional independence with kernels]

4. Kernel dimension reduction

1. Covariance operators on RKHS
2. Independence and dependence with kernels
3. Conditional independence with kernels
4. Kernel dimension reduction

Covariance on RKHS

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

(H_X, k_X), (H_Y, k_Y): RKHS with kernels on Ω_X and Ω_Y, resp.

Assume $E[k_X(X,X)] < \infty, E[k_Y(Y,Y)] < \infty$.

Cross-covariance operator: $\Sigma_{YX} : H_X \to H_Y$

$$\Sigma_{YX} \equiv E[\Phi_Y(Y) \otimes \Phi_X(X)] - m_Y \otimes m_X$$

$$= m_{P_{YX}} - m_{P_Y \otimes P_X} \qquad \in H_Y \otimes H_X$$

Proposition

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

for all $f \in H_X, g \in H_Y$

– Note: a linear map is a (1,1)-tensor.

– *c.f.* Euclidean case

$V_{YX} = \mathrm{E}[YX^T] - \mathrm{E}[Y]\mathrm{E}[X]^T$: covariance matrix

$$(b, V_{YX} a) = Cov[b^T Y, a^T X]$$

– Fact: Σ_{YX} is Hilbert Schmidt operator.

$$\|\Sigma_{YX}\|_{HS}^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|\langle \psi_j, \Sigma_{YX}\varphi_i\rangle\right|^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|\langle m_{(XY)} - m_X \otimes m_Y, \varphi_i \otimes \psi_j\rangle\right|^2$$

$$= \left\| m_{(XY)} - m_X \otimes m_Y \right\|_{H_X \otimes H_Y}^2$$

– Integral expression:

$$(\Sigma_{YX} f)(y) = \int \left(k_Y(y,Y) - E[k_Y(y,Y)]\right) f(X)\, dP(X,Y)$$

∵) Plug $g = k_Y(y,\cdot)$ in Proposition.

Characterization of Independence

- Independence and Cross-covariance operator

Theorem

If the product kernel $k_X k_Y$ is characteristic on Ω_X x Ω_Y, then

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

proof)

$$\Sigma_{XY} = O \quad \Leftrightarrow \quad m_{P_{XY}} = m_{P_X \otimes P_Y}$$

$$\Leftrightarrow \quad P_{XY} = P_X \otimes P_Y$$ (by characteristic assumption)

– *c.f.* for Gaussian variables

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

– *c.f.* Characteristic function

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad E_{XY}[e^{\sqrt{-1}(uX+vY)}] = E_X[e^{\sqrt{-1}uX}]E_Y[e^{\sqrt{-1}vY}]$$

- Intuition: High-order moments

Suppose X and Y are $\mathbf{R}$-valued, and $k(x,u)$ admits the expansion

$$k(x,u) = 1 + c_1 xu + c_2 x^2 u^2 + c_3 x^3 u^3 + \cdots$$

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

W.r.t. basis $1, u, u^2, u^3, \ldots$, the random variables on RKHS are expressed by

$$\Phi(X) = k(X,u) \sim (1, c_1 X, c_2 X^2, c_3 X^3, \ldots)^T$$

$$\Phi(Y) = k(Y,u) \sim (1, c_1 Y, c_2 Y^2, c_3 Y^3, \ldots)^T$$

$$\Sigma_{YX} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & c_1^2 Cov[Y,X] & c_1 c_2 Cov[Y,X^2] & c_1 c_3 Cov[Y^3,X] & \cdots \\ 0 & c_2 c_1 Cov[Y^2,X] & c_2^2 Cov[Y^2,X^2] & c_2 c_3 Cov[Y^2,X^3] & \cdots \\ 0 & c_3 c_1 Cov[Y^3,X] & c_3 c_2 Cov[Y^3,X^2] & c_3^2 Cov[Y^3,X^3] & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The operator Σ_{YX} contains all the high-order moments between X and Y.

Estimation of Cross-covariance Operator

$(X_1, Y_1), \ldots, (X_N, Y_N)$: i.i.d. sample on X x Y

An estimator of Σ_{YX} is defined by

$$\hat{\Sigma}_{YX}^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \left\{ k_{\mathrm{Y}}(\cdot, Y_i) - \hat{m}_Y \right\} \otimes \left\{ k_{\mathrm{X}}(\cdot, X_i) - \hat{m}_X \right\}$$

Theorem

$$\left\| \hat{\Sigma}_{YX}^{(N)} - \Sigma_{YX} \right\|_{HS} = O_p\left(1/\sqrt{N}\right) \qquad (N \to \infty)$$

Corollary to the $\sqrt{N}$-consistency of the empirical mean, because the norm in $H_{\mathrm{X}} \otimes H_{\mathrm{Y}}$ is equal to the Hilbert-Schmidt norm of the corresponding operator $H_{\mathrm{X}} \to H_{\mathrm{Y}}$

1. Covariance operators on RKHS
2. Independence and dependence with kernels
3. Conditional independence with RKHS
4. kernel dimension reduction

Measuring Dependence

- (In)dependence measure (HSIC, Hilbert-Schmidt Independence Criterion, Gretton et al 2005)

$$M_{YX} = \left\| \Sigma_{YX} \right\|_{HS}^2$$

$$M_{YX} = 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

with $k_X k_Y$ characteristic

- Empirical dependence measure

$$\hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2$$

M_{YX} and $\hat{M}_{YX}^{(N)}$ can be used as measures of dependence.

HS-norm of Cross-covariance Operator

- Empirical estimator

  Gram matrix expression

  HS-norm can be evaluated only in the subspaces $\mathrm{Span}\{k_{\mathcal{X}}(\cdot, X_i) - \hat{m}_X^{(N)}\}_{i=1}^N$ and $\mathrm{Span}\{k_{\mathcal{Y}}(\cdot, Y_i) - \hat{m}_Y^{(N)}\}$.

$$\Rightarrow \quad \hat{M}_{YX}^{(N)} = \frac{1}{N^2} \mathrm{Tr}[G_X G_Y]$$

where $G_X = Q_N K_X Q_N, \qquad Q_N = I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$

Or equivalently,

$$\hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 = \frac{1}{N^2} \sum_{i,j=1}^N k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{Y}}(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^N k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{Y}}(Y_i, Y_k) + \frac{1}{N^4} \sum_{i,j=1}^N k_{\mathcal{X}}(X_i, X_j) \sum_{k,\ell=1}^N k_{\mathcal{Y}}(Y_k, Y_\ell)$$

Normalized Covariance Operator

- Normalized Cross-Covariance Operator

NOCCO $W_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$

- Characterization of independence

With characteristic kernels,

$$W_{YX} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

Assume W_{XY} etc. are Hilbert-Schmidt.

– Dependence measure

$$\mathrm{NOCCO} = \|W_{YX}\|_{HS}^2$$

Kernel-free Integral Expression

Theorem (Fukumizu et al. NIPS 21, 2008)

Assume

P_{XY} have density $p_{XY}(x,y)$

$H_X \otimes H_Y$ are characteristic.

W_{YX} is Hilbert-Schmidt.

Then,

$$\| W_{YX} \|_{HS}^2 = \iint \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1 \right)^2 p_X(x) p_Y(y) dxdy$$

- Kernel-free expression, though the definitions are given by kernels!

- The RHS is χ^2-divergence (mean square contingency), which is a well-known dependence measure

Empirical Estimator

– Empirical estimation is straightforward with the empirical cross-covariance operator $\hat{\Sigma}_{YX}^{(N)}$.

– Inversion → regularization: $\Sigma_{XX}^{-1} \to \left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon I\right)^{-1}$

– Replace the covariances in $W_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ by the empirical ones given by the data $\Phi_X(X_1),\ldots,\Phi_X(X_N)$ and $\Phi_Y(Y_1),\ldots,\Phi_Y(Y_N)$

$$\mathrm{NOCCO}_{emp} = \mathrm{Tr}\left[R_X R_Y\right]$$ (dependence measure)

where $R_X \equiv G_X\left(G_X + N\varepsilon_N I_N\right)^{-1}$

$$G_X = \left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)K_X\left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right) \qquad K_X = \left(k(X_i, X_j)\right)_{i,j=1}^N$$

– NOCCO_{emp} gives a new kernel estimator for the χ^2-divergence. Consistency is known.

Independence Test with Kernels I

- Independence test with positive definite kernels
  - Null hypothesis H_0: X and Y are independent
  - Alternative H_1: X and Y are not independent

$\hat{M}_{YX}^{(N)}$ and NOCCO_{emp} can be used for test statistics.

$$\hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} k_X(X_i, X_j) k_Y(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^{N} k_X(X_i, X_j) k_Y(Y_i, Y_k) + \frac{1}{N^4} \sum_{i,j=1}^{N} k_X(X_i, X_j) \sum_{k,\ell=1}^{N} k_Y(Y_k, Y_\ell)$$

Independence test with kernels II

- Asymptotic distribution under null-hypothesis

Theorem (Gretton et al. 2008)

If X and Y are independent, then

$$N \hat{M}_{YX}^{(N)} \Rightarrow \sum_{i=1}^{\infty} \lambda_i Z_i^2 \quad \text{in law} \quad (N \to \infty)$$

where

Z_i : i.i.d. ~ $N(0,1)$,

$\{\lambda_i\}_{i=1}^{\infty}$ is the eigenvalues of the following integral operator

$$\int h(u_a, u_b, u_c, u_d) \varphi_i(u_b) dP_{U_b} dP_{U_c} dP_{U_d} = \lambda_i \varphi_i(u_a)$$

$$h(U_a, U_b, U_c, U_d) = \tfrac{1}{4!} \sum_{(a,b,c,d)} k_{a,b}^{\mathrm{X}} k_{a,b}^{\mathrm{Y}} - 2 k_{a,b}^{\mathrm{X}} k_{a,c}^{\mathrm{Y}} + k_{a,b}^{\mathrm{X}} k_{c,d}^{\mathrm{Y}}$$

$$k_{a,b}^{\mathrm{X}} = k_{\mathrm{X}}(X_a, X_b), \quad U_a = (X_a, Y_a)$$

– The proof is standard by the theory of degenerate U (or V)-statistics (see e.g. Serfling 1980, Chapter 5).

Independence test with kernels III

- Consistency of test

<u>Theorem (Gretton et al. 2008)</u>

If M_{YX} is not zero, then

$$\sqrt{N}\left(\hat{M}_{YX}^{(N)} - M_{YX}\right) \;\Rightarrow\; N(0,\sigma^2) \quad \text{in law} \quad (N \to \infty)$$

where

$$\sigma^2 = 16\left(E_a\left[E_{b,c,d}[h(U_a,U_b,U_c,U_d)]^2\right] - M_{YX}\right)$$

Choice of Kernel

- How to choose a kernel?
  - No definitive solutions have been proposed yet.
  - For statistical tests, comparison of power or efficiency will be desirable.
  - Other suggestions:
    - Make a relevant supervised problem, and use cross-validation.
    - Some heuristics
      - Heuristics for Gaussian kernels (Gretton et al 2007)

$$\sigma = \text{median}\left\{ \left\| X_i - X_j \right\| \mid i \neq j \right\}$$

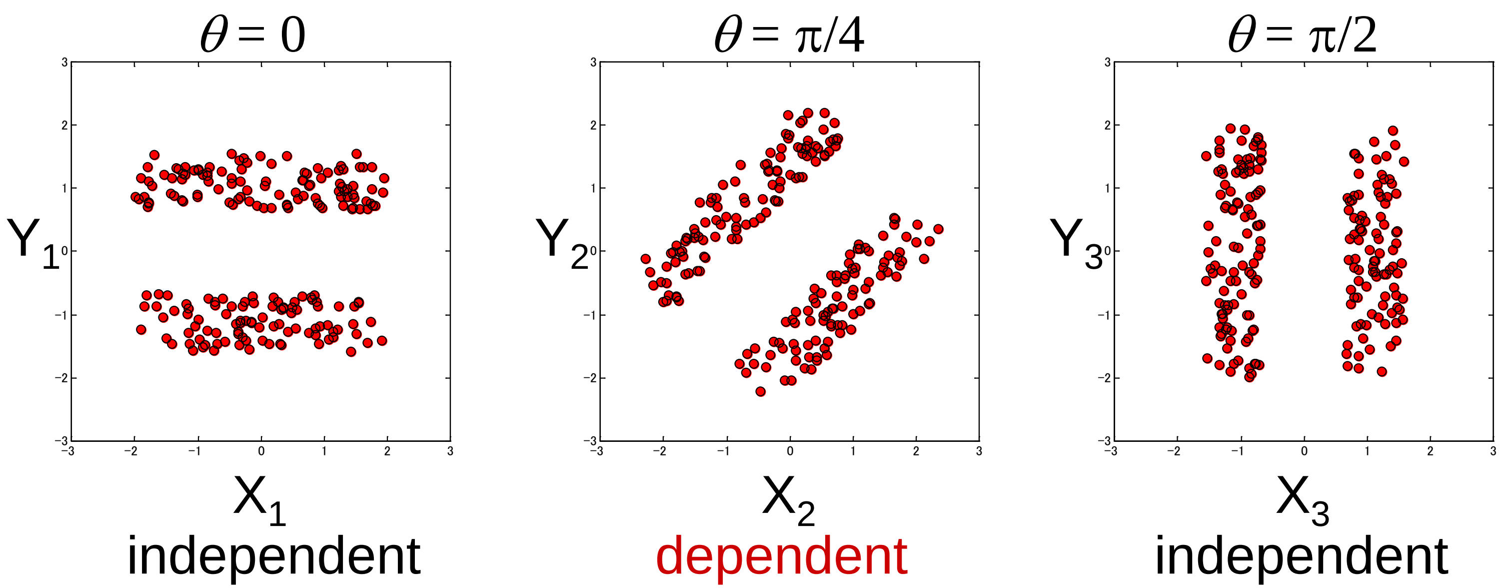
      - Speed of asymptotic convergence (Fukumizu et al. 2008)

$$\lim_{N \to \infty} Var\left[N \times HSIC_{emp}^{(N)} \right] = 2 \left\| \Sigma_{XX} \right\|_{HS}^2 \left\| \Sigma_{YY} \right\|_{HS}^2 \quad \text{under independence}$$

Compare the bootstrapped variance and the theoretical one, and choose the parameter to give the minimum discrepancy.

Application to Independence Test

- Toy example



They are all uncorrelated, but dependent for $0 < \theta < \pi/2$

N = 200.

Permutation test is used for independence test except contingency table.

indep. ⟶ more dependent

Angle	0.0	4.5	9.0	13.5	18.0	22.5
HSIC (Median)	93	92	63	5	0	0
HSIC (Asymp. Var.)	93	44	1	0	0	0
NOCCO ($\varepsilon = 10^4$, Median)	94	**23**	**0**	**0**	**0**	**0**
NOCCO ($\varepsilon = 10^6$, Median)	92	**20**	**1**	**0**	**0**	**0**
NOCCO ($\varepsilon = 10^8$, Median)	93	**15**	**0**	**0**	**0**	**0**
NOCCO (Asymp. Var.)	94	**11**	**0**	**0**	**0**	**0**
MI (#NN = 1)	93	62	11	0	0	0
MI (#NN = 3)	96	43	0	0	0	0
MI (#NN = 5)	97	49	0	0	0	0
Power Diverg. (#Bins=3)	96	92	43	9	1	0
Power Diverg. (#Bins=4)	98	29	0	0	0	0
Power Diverg. (#Bins=5)	94	60	2	0	0	0

acceptance of independence out of 100 tests (significance level = 5%)

MI: mutual information estimated by the nearest neighbor method.

- Power Divergence (Ku&Fine05, Read&Cressie)
  - Make partition $\{A_j\}_{j\in J}$: Each dimension is divided into q parts so that each bin contains almost the same number of data.
  - Power-divergence

$$T_N = 2I^{\lambda}(X,m) = N\frac{2}{\lambda(\lambda+2)}\sum_{j\in J}\hat{p}_j\left\{\left(\hat{p}_j \Big/ \prod_{k=1}^{N}\hat{p}^{(k)}_{j_k}\right)^{\lambda} - 1\right\}$$

I^0 = MI

I^2 = Mean Square Conting.

$\hat{p}_j$: frequency in A_j

$\hat{p}^{(k)}_r$: marginal freq. in r-th interval

  - Null distribution under independence

$$T_N \quad \Rightarrow \quad \chi^2_{q^N - qN + N - 1}$$

Independent Test on Text

- Data: Official records of Canadian Parliament in English and French.
  - Dependent data: 5 line-long parts from English texts and their French translations.
  - Independent data: 5 line-long parts from English texts and random 5 line-parts from French texts.
- Kernel: Bag-of-words and spectral kernel

Results of permutations test with HS measure

Topic	Match	BOW(N=10)	Spec(N=10)	BOW(N=50)	Spec(N=50)
Agri-culture	Random	0.94	0.95	0.93	0.95
	Same	0.18	0.00	0.00	0.00
Fishery	Random	0.94	0.94	0.93	0.95
	Same	0.20	0.00	0.00	0.00
Immig-ration	Random	0.96	0.91	0.94	0.95
	Same	0.09	0.00	0.00	0.00

Acceptance rate ($\alpha = 5\%$)

(Gretton et al. 2007)

Independence Test: Comparison

– Brownian distance covariance (Székely & Rizzo AOAS 2010)

- Independence with the characteristic functions

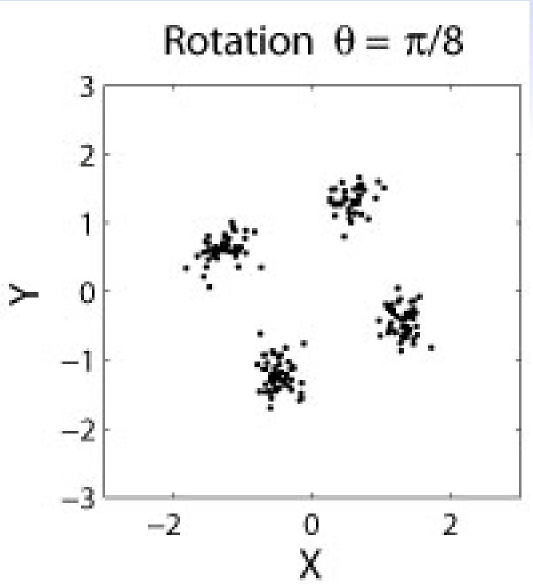
$$X \perp\!\!\!\perp Y \qquad \Leftrightarrow \qquad \phi_{XY} = \phi_X \phi_Y$$

$$\phi_X(\omega) = E[e^{\sqrt{-1}X^T\omega}], \;\; \phi_Y(\xi) = E[e^{\sqrt{-1}Y^T\xi}], \; \phi_{XY}(\omega, \xi) = E[e^{\sqrt{-1}(X^T\omega + Y^T\xi)}].$$

- Independence measure with weighted integral:

$$\int \left| \phi_{XY}(\omega, \xi) - \phi_X(\omega)\phi_Y(\xi) \right|^2 w(\omega, \xi) d\omega d\xi$$

- With a clever choice of the weight w, the integral is reduced to HSIC-like measure with $k(x_1, x_2) = \|x_1 - x_2\|$



indep. ⟶ more dependent

angle :		0	π/12	π/6	π/4
$d_X = d_Y = 2$	HS	0.94	0.77	0.48	0.42
$N = 128$	SR	0.95	0.83	0.66	0.65
$d_X = d_Y = 2$	HS	0.92	0.47	0.17	0.12
$N = 512$	SR	0.93	0.49	0.38	0.33
$d_X = d_Y = 4$	HS	0.92	0.60	0.35	0.23
$N = 1024$	SR	0.93	0.68	0.48	0.47
$d_X = d_Y = 4$	HS	0.92	0.44	0.15	0.12
$N = 2048$	SR	0.94	0.46	0.29	0.27

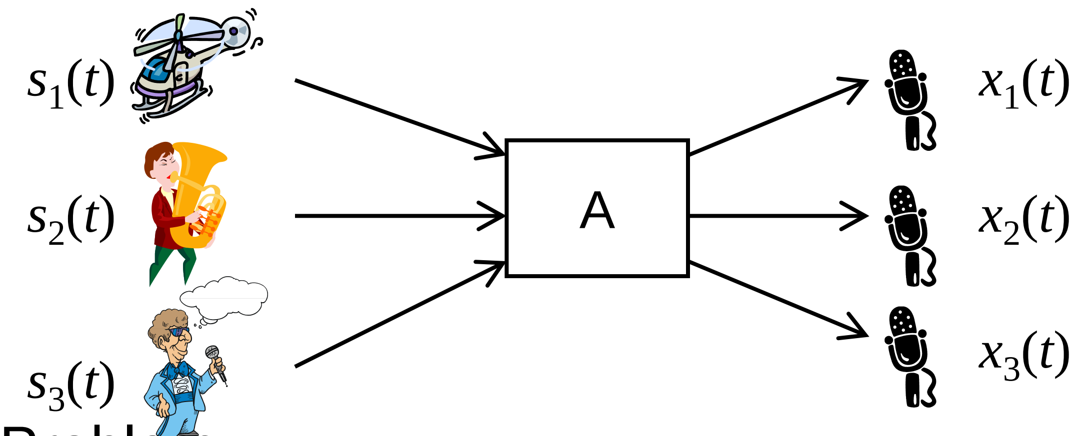
% of acceptance of indep. in permutation tests (α = 5%).

HS: Hilbert-Schmidt norm.
Gaussian kernel
$\sigma = \mathrm{med}\{\|X_i - X_j\|\}$
SR: Székely & Rizzo

(Gretton, F, Sriperumbudur. 2010. AOAS Discussion)

Application: ICA

- Independent Component Analysis (ICA)
  - Assumption
    - m independent source signals
    - m observations of linearly mixed signals



$$X(t) = AS(t)$$

A: m x m invertible matrix

  - Problem
    - Restore the independent signals S from observations X.

$$\hat{S} = BX$$

B: m x m orthogonal matrix

- ICA with HS independence measure

$X^{(1)},...,X^{(N)}$: i.i.d. observation (m-dimensional)

Pairwise-independence criterion is applicable.

Minimize $$L(B) = \sum_{a=1}^{m} \sum_{b>a} \hat{M}(Y_a, Y_b) \qquad Y = BX$$

Objective function is non-convex. Optimization is not easy.

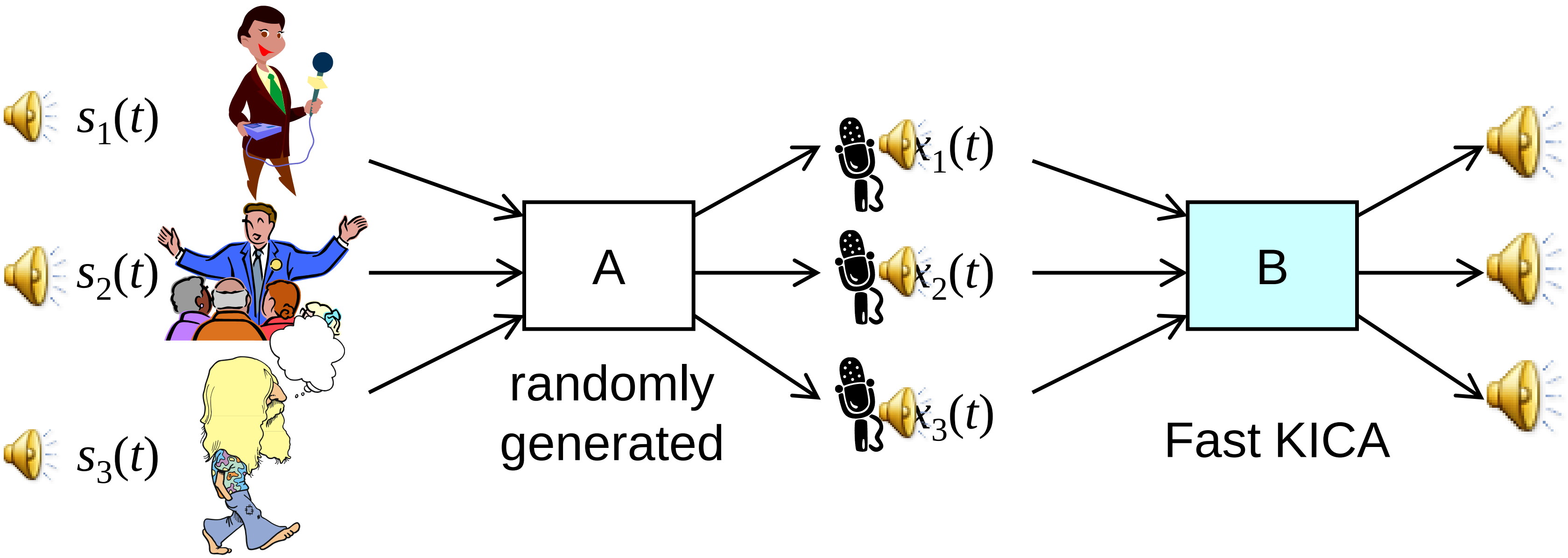
→ Approximate Newton method has been proposed

Fast Kernel ICA (FastKICA, Shen et al 07)
(Software downloadable at Arthur Gretton's homepage)

- Other methods for ICA

See, for example, Hyvärinen et al. (2001).

- Experiments (speech signal)

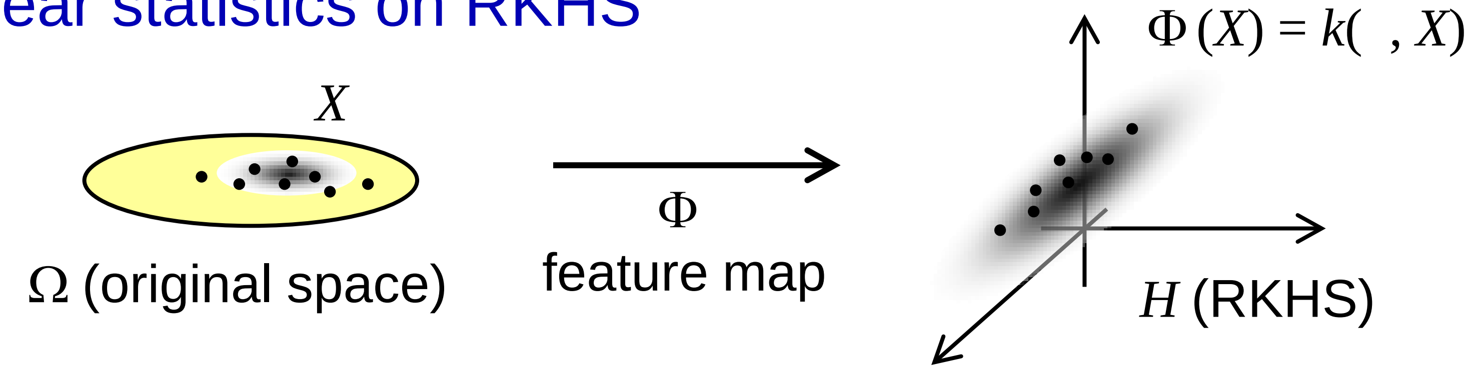


Three speech signals

1. Covariance operators on RKHS
2. Independence and dependence with kernels
3. Conditional independence with kernels
4. Kernel dimension reduction

Re: Statistics on RKHS

- Linear statistics on RKHS



– Basic statistics on Euclidean space → Basic statistics on RKHS

Basic statistics on Euclidean space		Basic statistics on RKHS
Mean	$\longrightarrow$	Kernel mean
Covariance	$\longrightarrow$	Cross-covariance operator Σ_{YX}
Conditional covariance	$\longrightarrow$	Cond. cross-covariance operator

– Plan: define the basic statistics on RKHS and derive nonlinear/ nonparametric statistical methods in the original space.

Conditional Independence

- Definition

  X, Y, Z: random variables with joint p.d.f. $p_{XYZ}(x, y, z)$

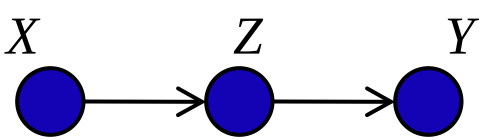
  X and Y are conditionally independent given Z, if

$$p_{Y|ZX}(y \mid z, x) = p_{Y|Z}(y \mid z) \qquad \text{(A)}$$

or

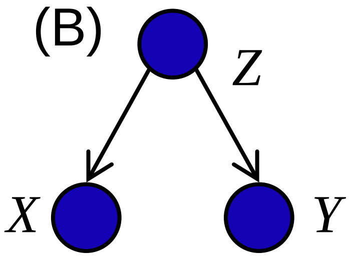
$$p_{XY|Z}(x, y \mid z) = p_{X|Z}(x \mid z)\, p_{Y|Z}(y \mid z) \qquad \text{(B)}$$

(A)



With Z known, the information of X is unnecessary for the inference on Y

(B)



- Applications
  - Graphical model
  - Causal inference, etc.

Conditional Independence for Gaussian Variables

- Two characterizations

  X, Y, Z are Gaussian.

  - Conditional covariance

    $$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{XY|Z} = O \qquad \text{i.e.} \quad V_{YX} - V_{YZ} V_{ZZ}^{-1} V_{ZX} = O$$

  - Comparison of conditional variance

    $$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{YY|[X,Z]} = V_{YY|Z}$$

Linear Regression and Conditional Covariance

- Review: linear regression
  - X, Y: random vector (not necessarily Gaussian) of dim p and q.
    $\tilde{X} = X - E[X], \quad \tilde{Y} = Y - E[Y]$
  - Linear regression: predict Y using the linear combination of X.
    Minimize the mean square error:
    $$\min_{A: q\times p \text{ matrix}} E\left\|\tilde{Y} - A\tilde{X}\right\|^2$$
  - The residual error is given by the conditional covariance matrix.
    $$\min_{A: q\times p \text{ matrix}} E\left\|\tilde{Y} - A\tilde{X}\right\|^2 = \mathrm{Tr}\left[V_{YY|X}\right]$$
  - For Gaussian variables, $V_{YY|[X,Z]} = V_{YY|Z} \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$
    can be interpreted as
    "If Z is known, X is not necessary for linear prediction of Y."

Review: Conditional Covariance

- Conditional covariance of Gaussian variables
  - Jointly Gaussian variable

    $$X = (X_1, \ldots, X_p), Y = (Y_1, \ldots, Y_q)$$

    $Z = (X, Y)$: m $(= p + q)$ dimensional Gaussian variable

    $$Z \sim N(\mu, V) \qquad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \qquad V = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix}$$

  - Conditional probability of Y given X is again Gaussian

    $$\sim N(\mu_{Y|X}, V_{YY|X})$$

    Cond. mean $\quad \mu_{Y|X} \equiv E[Y \mid X = x] = \mu_Y + V_{YX} V_{XX}^{-1} (x - \mu_X)$

    Cond. covariance $\quad V_{YY|X} \equiv Var[Y \mid X = x] = \underline{V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}}$

    Schur complement of V_{XX} in V

    Note: $V_{YY|X}$ does not depend on x

Conditional Covariance on RKHS

- Conditional Cross-covariance operator

  X, Y, Z : random variables on Ω_X, Ω_Y, Ω_Z (resp.).

  (H_X, k_X), (H_Y, k_Y), (H_Z, k_Z) : RKHS defined on Ω_X, Ω_Y, Ω_Z (resp.).

  - Conditional cross-covariance operator

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

  - Conditional covariance operator

$$\Sigma_{YY|Z} \equiv \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} \quad : \quad H_Y \to H_Y$$

  - Σ_{ZZ}^{-1} may not exist as a bounded operator. But, we can justify the definitions.

- Decomposition of covariance operator

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2}$$

such that W_{YX} is a bounded operator with $\| W_{YX} \| \le 1$ and

$$\overline{Range(W_{YX})} = \overline{Range(\Sigma_{YY})}, \quad Ker(W_{YX}) \perp \overline{Range(\Sigma_{XX})}.$$

W_{YX} is the ‘correlation’ operator.
$\Sigma_{XX}^{1/2}$ is defined by the eigendecomposition.

- Rigorous definitions

$$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZX} \Sigma_{XX}^{1/2}$$

$$\Sigma_{YY|Z} \equiv \Sigma_{YY} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZY} \Sigma_{YY}^{1/2}$$

Conditional Covariance

- Conditional covariance is expressed by operators

Proposition (FBJ 2004, 2008+)

Assume k_Z is characteristic.

$$\langle g, \Sigma_{YX|Z} f \rangle = E\big[\mathrm{Cov}[g(Y), f(X) \mid Z]\big] \qquad (\forall f \in H_X, g \in H_Y)$$

In particular,

$$\langle g, \Sigma_{YY|Z} g \rangle = E\big[\mathrm{Var}[g(Y) \mid Z]\big] \qquad (\forall g \in H_Y)$$

Proof omitted.

Analogy to Gaussian variables:

$$b^T (V_{YX} - V_{YZ} V_{ZZ}^{-1} V_{ZX}) a = \mathrm{Cov}[b^T Y, a^T X \mid Z]$$

$$b^T (V_{YY} - V_{YZ} V_{ZZ}^{-1} V_{ZY}) b = \mathrm{Var}[b^T Y \mid Z]$$

Mean Square Error Interpretation

<u>Proposition</u> <u>(FBJ 2004, 2009)</u>

Assume k_Z is characteristic.

$$\left\langle g, \Sigma_{YY|Z} g \right\rangle = E\big[Var[g(Y) \mid Z]\big] = \inf_{f \in H_Z} E\left|\tilde{g}(Y) - \tilde{f}(Z)\right|^2 \qquad (\forall g \in H_Y)$$

where $\tilde{f}(X) = f(X) - E[f(X)], \ \tilde{g}(Y) = g(Y) - E[g(Y)].$

c.f. for Gaussian variables

$$b^T V_{YY|Z} b = Var[b^T Y \mid Z] = \min_a \left| b^T \tilde{Y} - a^T \tilde{Z} \right|^2$$

– Proof (left = right)

$$E\left|\left(g(Y)-E[g(Y)]\right)-\left(f(Z)-E[f(Z)]\right)\right|^2$$

$$=\langle f,\Sigma_{ZZ}f\rangle-2\langle f,\Sigma_{ZY}g\rangle+\langle g,\Sigma_{YY}g\rangle$$

$$=\left\|\Sigma_{ZZ}^{1/2}f\right\|^2-2\left\langle f,\Sigma_{ZZ}^{1/2}W_{ZY}\Sigma_{YY}^{1/2}g\right\rangle+\left\|\Sigma_{YY}^{1/2}g\right\|^2$$

$$=\left\|\Sigma_{ZZ}^{1/2}f-W_{ZY}\Sigma_{YY}^{1/2}g\right\|^2+\left\|\Sigma_{YY}^{1/2}g\right\|^2-\left\|W_{ZY}\Sigma_{YY}^{1/2}g\right\|^2$$

$$=\underline{\left\|\Sigma_{ZZ}^{1/2}f-W_{ZY}\Sigma_{YY}^{1/2}g\right\|^2}+\left\langle g,\underline{\left(\Sigma_{YY}-\Sigma_{YY}^{1/2}W_{YZ}W_{ZY}\Sigma_{YY}^{1/2}\right)}g\right\rangle$$

This part can be arbitrary small by choosing *f* because of $\overline{Range(W_{ZY})}=\overline{Range(\Sigma_{ZZ})}$.

$\Sigma_{YY|Z}$

Conditional Independence with Kernels

Theorem (FBJ2004, 2008+)

Assume k_Z and $k_X k_Y k_Z$ are characteristic.

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad \Sigma_{\ddot{Y}X|Z} = O$$

where $\ddot{Y} = (Y, Z)$

Assume k_Z , k_Y, $k_X k_Z$ are characteristic.

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad \Sigma_{YY|[X\ Z]} = \Sigma_{YY|Z}$$

– *c.f.* Gaussian variables

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{XY|Z} = O$$

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{YY|[X,Z]} = V_{YY|Z}$$

– Intuition of the condition $\Sigma_{YY|[X\ Z]} = \Sigma_{YY|Z}$

$$\text{Var}[g(Y) \mid X, Z] = \text{Var}[g(Y) \mid X]$$

If we already know X, the mean square error in predicting Y does not decrease, if information Z is added.

In general, $\Sigma_{YY|[X\ Z]} \le \Sigma_{YY|Z}$.

Empirical Estimator of Conditional Covariance Operator

$(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)$

$\Sigma_{YZ} \rightarrow \hat{\Sigma}_{YZ}^{(N)}$ etc. finite rank operators

$\Sigma_{ZZ}^{-1} \rightarrow \left(\hat{\Sigma}_{ZZ}^{(N)} + \varepsilon_N I\right)^{-1}$ regularization for inversion

– Empirical conditional covariance operator

$$\hat{\Sigma}_{YX|Z}^{(N)} := \hat{\Sigma}_{YX}^{(N)} - \hat{\Sigma}_{YZ}^{(N)}\left(\hat{\Sigma}_{ZZ}^{(N)} + \varepsilon_N I\right)^{-1}\hat{\Sigma}_{ZX}^{(N)}$$

– Estimator of Hilbert-Schmidt norm

$$\left\|\hat{\Sigma}_{YX|Z}^{(N)}\right\|_{HS}^2 = \mathrm{Tr}\left[G_X S_Z G_Y S_Z\right]$$

$$G_X = Q_N K_X Q_N, \quad Q_N = I_N - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^T$$ centered Gram matrix

$$S_Z = I_N - (G_Z + N\varepsilon_N I_N)^{-1} G_Z = \left(I_N + \tfrac{1}{N\varepsilon_N} G_Z\right)^{-1}$$

Consistency

- Consistency on conditional covariance operator

Theorem (FBJ08, Sun et al. 07)

Assume $\varepsilon_N \to 0$ and $\sqrt{N}\varepsilon_N \to \infty$

$$\left\| \hat{\Sigma}_{YX|Z}^{(N)} - \Sigma_{YX|Z} \right\|_{HS} \to 0 \qquad (N \to \infty)$$

In particular,

$$\left\| \hat{\Sigma}_{YX|Z}^{(N)} \right\|_{HS} \to \left\| \Sigma_{YX|Z} \right\|_{HS} \qquad (N \to \infty)$$

Applications of Conditional Independence

- Conditional independence test (Fukumizu et al. 2008)
  - Estimation of graphical model by data.

- Causality
  - Causal relations among variables can be formulated in terms of conditional independence or Markov network. (Sun et al 2007)
  - Granger causality for time series.

    (X_t) is not a cause of (Y_t) if

$$p(Y_t \mid Y_{t-1},\dots,Y_{t-p}, X_{t-1},\dots,X_{t-p}) = p(Y_t \mid Y_{t-1},\dots,Y_{t-p})$$

$$\Leftrightarrow \quad Y_t \perp\!\!\!\perp X_{t-1},\dots,X_{t-p} \mid Y_{t-1},\dots,Y_{t-p}$$

- And more

1. Covariance operators on RKHS
2. Independence and dependence with kernels
3. Conditional independence with kernels
4. Kernel dimension reduction

Dimension Reduction for Regression

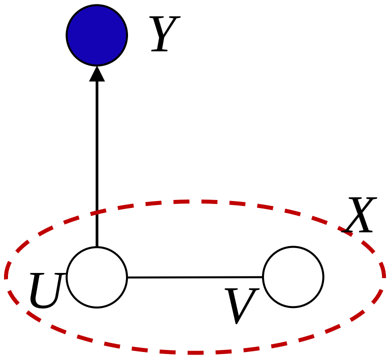
– Regression: Y : response variable,
$X=(X_1,...,X_m)$: m-dim. explanatory variable

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

$$p(Y \mid X) = \tilde{p}(Y \mid b_1^T X, ..., b_d^T X) \quad \left(= \tilde{p}(Y \mid B^T X) \right)$$

$B=(b_1,..,b_d)$: $m \times d$ matrix d is fixed.

$$\Longleftrightarrow \quad X \perp\!\!\!\perp Y \mid B^T X$$



$U = B^T X$

45

Kernel Dimension Reduction

(Fukumizu, Bach, Jordan JMLR 2004, AS 2009)

Use characteristic kernels for $B^T X$ and Y.

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

$$\Sigma_{YY|B^T X} = \Sigma_{YY|X} \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid B^T X$$ EDR space

– KDR objective function

$$\min_{B:\, B^T B = I_d} \operatorname{Tr}\left[\Sigma_{YY|B^T X} \right]$$

Equivalently

$$\min_{B: B^T B = I_d} \sum_{i=1}^{\infty} \inf E \left| \left(\psi_i(Y) - E[\psi_i(Y)] \right) - \left(f(B^T X) - E[f(B^T X)] \right) \right|^2$$

$\{\psi_i\}_{i=1}^{\infty}$: ONB of H_Y

– KDR empirical objective function

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

KDR method

- Wide applicability of KDR
  - The most general approach to dimension reduction:
    - no strong 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, such as the elliptic assumption for $p(X)$ for SIR.

- Computational issues
  - Non-convex objective function, possibly local minima.
    → Gradient method with an annealing technique
    starting from a large σ in Gaussian RBF kernel.
  - Computational cost with matrices of sample size.
    → Low-rank approximation.

47

Consistency of KDR

<u>Theorem (FBJ2009)</u>

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;

$$S_0 = \left\{ B \mid B^T B = I_d, \operatorname{Tr}\left[\Sigma_{YY|B^T X}\right] = \min_{B'} \operatorname{Tr}\left[\Sigma_{YY|B'^T X}\right] \right\}$$

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

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

$$\Pr\left(\hat{B}^{(N)} \in U \right) \to 1 \;\;\; (N \to \infty).$$

48

Numerical Results with KDR

- Synthetic data

X : 4 dim. $\sim N(0, I_4)$

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

Sample size N = 100

	KDR		SIR		SAVE		pHd	
τ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.1	0.11	±0.07	0.55	±0.28	0.77	±0.35	1.04	±0.34
0.4	0.17	±0.09	0.60	±0.27	0.82	±0.34	1.03	±0.33
0.8	0.34	±0.22	0.69	±0.25	0.94	±0.35	1.06	±0.33

Frobenius norms between the estimator and the true one
over 100 samples (means and standard deviations)

49

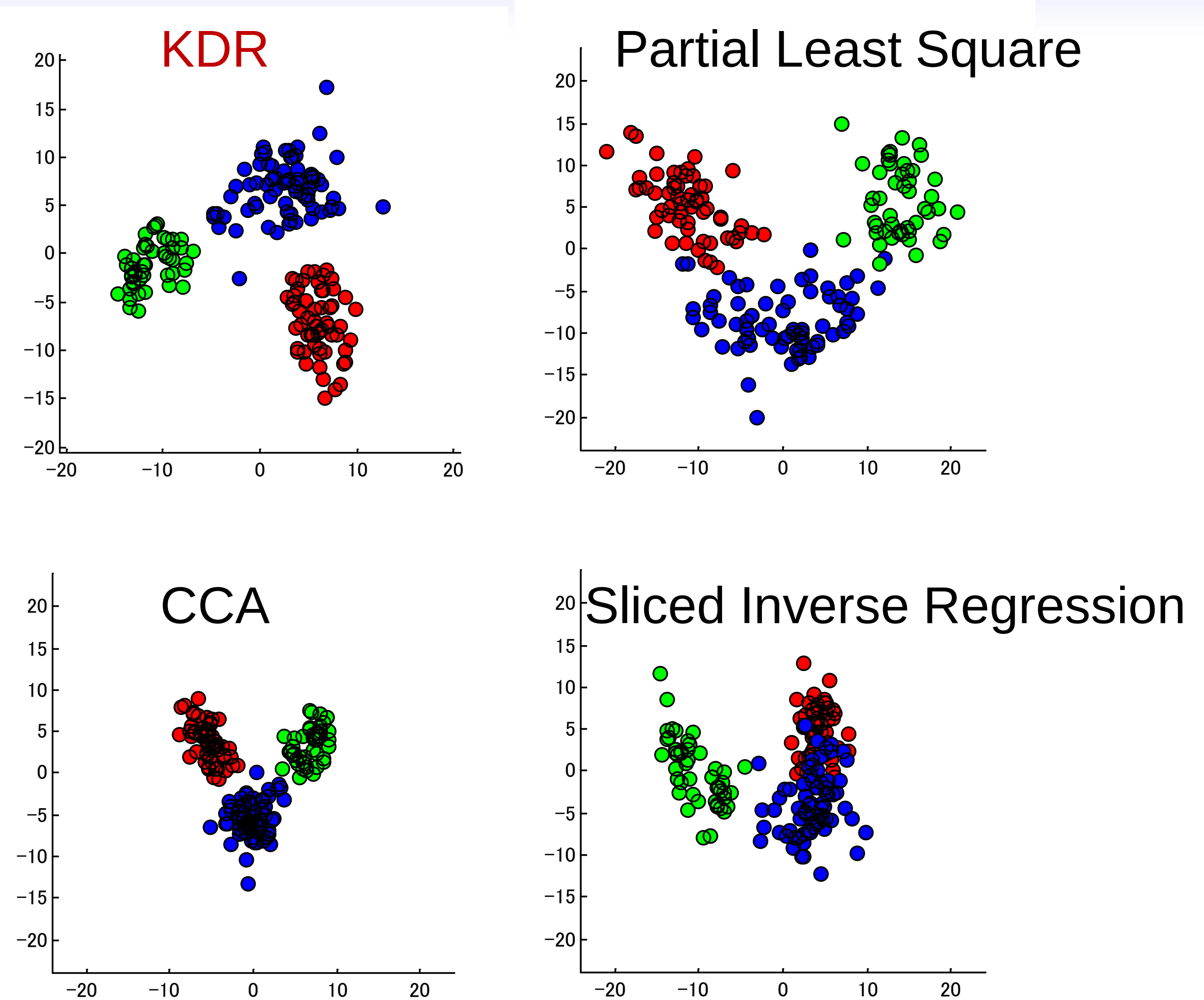
- Wine data

  13 dim. 178 data.

  Y = 3 class label

  2 dim. projection

$$k(z_1, z_2) = \exp\left(-\|z_1 - z_2\|^2 / \sigma^2\right)$$



50

Summary

- Dependence analysis with RKHS
  - Covariance and conditional covariance on RKHS can capture the (in)dependence and conditional (in)dependence of random variables.
  - Easy estimators can be obtained for the Hilbert-Schmidt norm of the operators.
  - If the normalized covariance is used, the Hilbert-Schmidt norm is independent of kernel (χ^2-divergence), assuming it is characteristic.
  - Statistical tests of independence and conditional independence are possible with kernel measures.
  - Applications: dimension reduction for regression (FBJ04, FBJ09), causal inference (Sun et al. 2007).

References

Fukumizu, K. Francis R. Bach and M. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*. 37(4), pp.1871-1905 (2009).

Fukumizu, K., A. Gretton, X. Sun, and B. Schölkopf: Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 21*, 489-496, MIT Press (2008).

Fukumizu, K., Bach, F.R., and Jordan, M.I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*. 5(Jan):73-99, 2004.

Gretton, A., K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and Alex Smola. A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems 20.* 585-592. MIT Press 2008.

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. 2007.

Gretton, A., O. Bousquet, A. Smola and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. Proc. Algorithmic Learning Theory (ALT2005), 63-78. 2005.

Shen, H., S. Jegelka and A. Gretton: Fast Kernel ICA using an Approximate Newton Method. AISTATS 2007.

Serfling, R. J. *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience 1980.

Sun, X., Janzing, D. Schölkopf, B., and Fukumizu, K.: A kernel-based causal learning algorithm. *Proc. 24th Annual International Conference on Machine Learning (ICML2007),* 855-862 (2007)