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

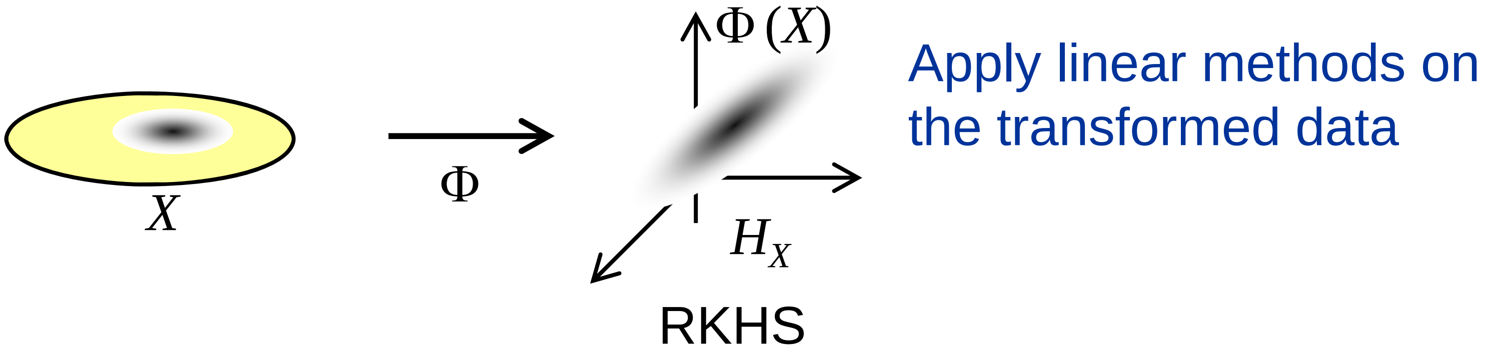
July 14-19, 2008. Singapore

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.

Positive definite kernel and RKHS

Positive definite kernel

Ω: set. $k : \Omega \times \Omega \to \mathbf{R}$

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

$\Rightarrow$ $\exists 1\ \mathcal{H}$: Hilbert space consisting of functions on Ω such that

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

2) $\mathrm{Span}\{k(\cdot, x) \mid x \in \Omega\}$ is dense in $\mathcal{H}$.

3) $\langle k(\cdot, x), f \rangle_{\mathcal{H}} = f(x)$ $\forall f \in \mathcal{H}, x \in \Omega$. (reproducing property)

How to use RKHS for data analysis?

Transform data into RKHS.

$$\Phi : \Omega \to \mathcal{H}, \quad x \mapsto k(\cdot, x)$$

$$i.e. \quad \Phi(x) = k(\cdot, x)$$

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

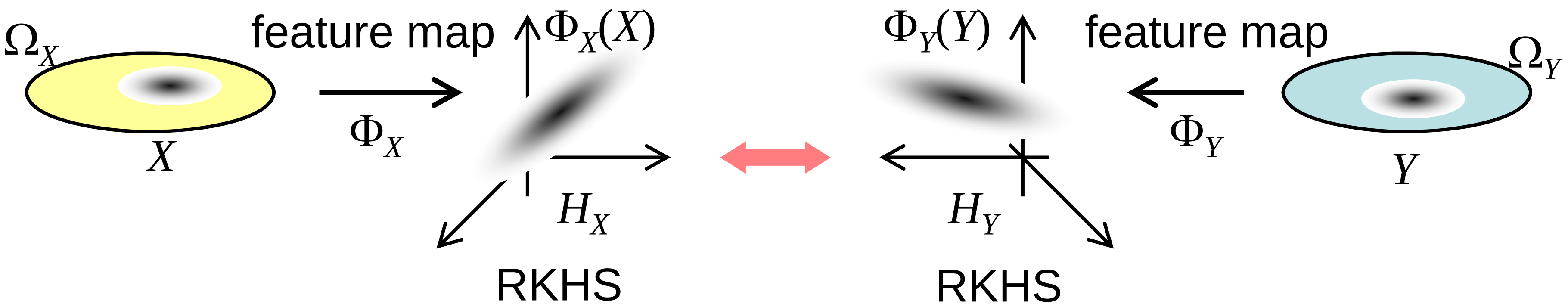


Illustration of dependence analysis with RKHS

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 Ω_X x Ω_Y.

$(\mathcal{H}_X, k_X)$, $(\mathcal{H}_Y, k_Y)$: RKHS on Ω_X and Ω_Y, resp.

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

$$\Phi_X(X) = k_X(\cdot, X), \qquad \Phi_Y(Y) = k_Y(\cdot, Y).$$

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

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

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

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

c.f. ordinary covariance matrix: $V_{XY} = \mathrm{Cov}[X, Y] = E[YX^T] - E[Y]E[X]^T$

Characterization of independence

Independence and cross-covariance operator

If the RKHS's are "rich enough" to express all the moments,

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad \Sigma_{XY} = O$$

$$\Leftrightarrow \quad E[g(Y)f(X)] = E[g(Y)]E[f(X)]$$

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 $P_X P_Y$.

– Analog to Gaussian random vectors: $X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad V_{YX} = O$.

– *c.f.* characteristic function

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad E_{XY}\left[e^{\sqrt{-1}\omega^T X} e^{\sqrt{-1}\eta^T Y}\right] = E_X\left[e^{\sqrt{-1}\omega^T X}\right] E_Y\left[e^{\sqrt{-1}\eta^T Y}\right]$$

for all ω and η.

– Applied to independence test (Gretton et al. 2008).

Characteristic kernels

- A class for determining a probability

X: random variable taking values on Ω.

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

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

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

($\mathcal{H}$ 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,

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad \Sigma_{XY} = O.$$

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

– An example on $\mathbf{R}^m$: 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 Ω_X x Ω_Y.

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

(rank $\le N$)

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

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

<u>Theorem</u>

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

– A uniform law of large numbers follows:

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

– Weak convergence of $\sqrt{N}\left(\hat{\Sigma}_{YX}^{(N)} - \Sigma_{YX}\right)$ 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. $\Sigma_{YY|X} \equiv \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$: 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),

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

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

– Empirical estimator

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

ε_N: regularization coefficient

Can be represented by Gram matrices.

Conditional independence

Theorem (FBJ 2004, 2006)

U, V, and Y are random variables on Ω_U, Ω_V, and Ω_Y, resp.

$\mathcal{H}_U$, $\mathcal{H}_V$, $\mathcal{H}_Y$: RKHS on Ω_U, Ω_V, Ω_Y with kernel k_U, k_V, k_Y, resp.

$X = (U,V)$. RKHS on $\Omega_X = \Omega_U$ x Ω_V is defined by $k_X = k_U k_V$.

Assume $\mathcal{H}_X$, $\mathcal{H}_U$: characteristic. Then,

$$\Sigma_{YY|U} \geq \Sigma_{YY|X}$$

$\geq$: the partial order of self-adjoint operators

If further $\mathcal{H}_Y$ is characteristic, then

$$Y \perp\!\!\!\perp X \mid U \quad \Leftrightarrow \quad \Sigma_{YY|U} = \Sigma_{YY|X}$$

$\mathrm{Tr}\left[\Sigma_{YY|U} - \Sigma_{YY|X}\right]$ 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)

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

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

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

– 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 $B^T X$ 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

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

– KDR contrast function with finite sample

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

where

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

$$K_{B^T X, ij} = k_d(B^T X_i, B^T X_j)$$

KDR method

- **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.
    → Low-rank approximation, *e.g*. incomplete Cholesky decomposition.
  - Non-convex contrast function, possibly local minima.
    → Gradient method with an annealing technique starting from a large σ in Gaussian RBF kernel.

Consistency of KDR

<u>Theorem (FBJ2006)</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;

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

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

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

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

Numerical results with KDR

- Synthetic data (A)

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 of the projection matrices over 100 samples.
(Means and standard deviations)

Synthetic data (B)

X : 10 dim. $\sim N(0, I_4)$

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

Sample size N = 500

	KDR		SIR		SAVE		pHd	
a	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.0	0.17	±0.05	1.83	±0.22	0.30	±0.07	1.48	±0.27
0.5	0.17	±0.04	0.58	±0.19	0.35	±0.08	1.52	±0.28
1.0	0.18	±0.05	0.30	±0.08	0.57	±0.20	1.58	±0.28

KDR on Real data

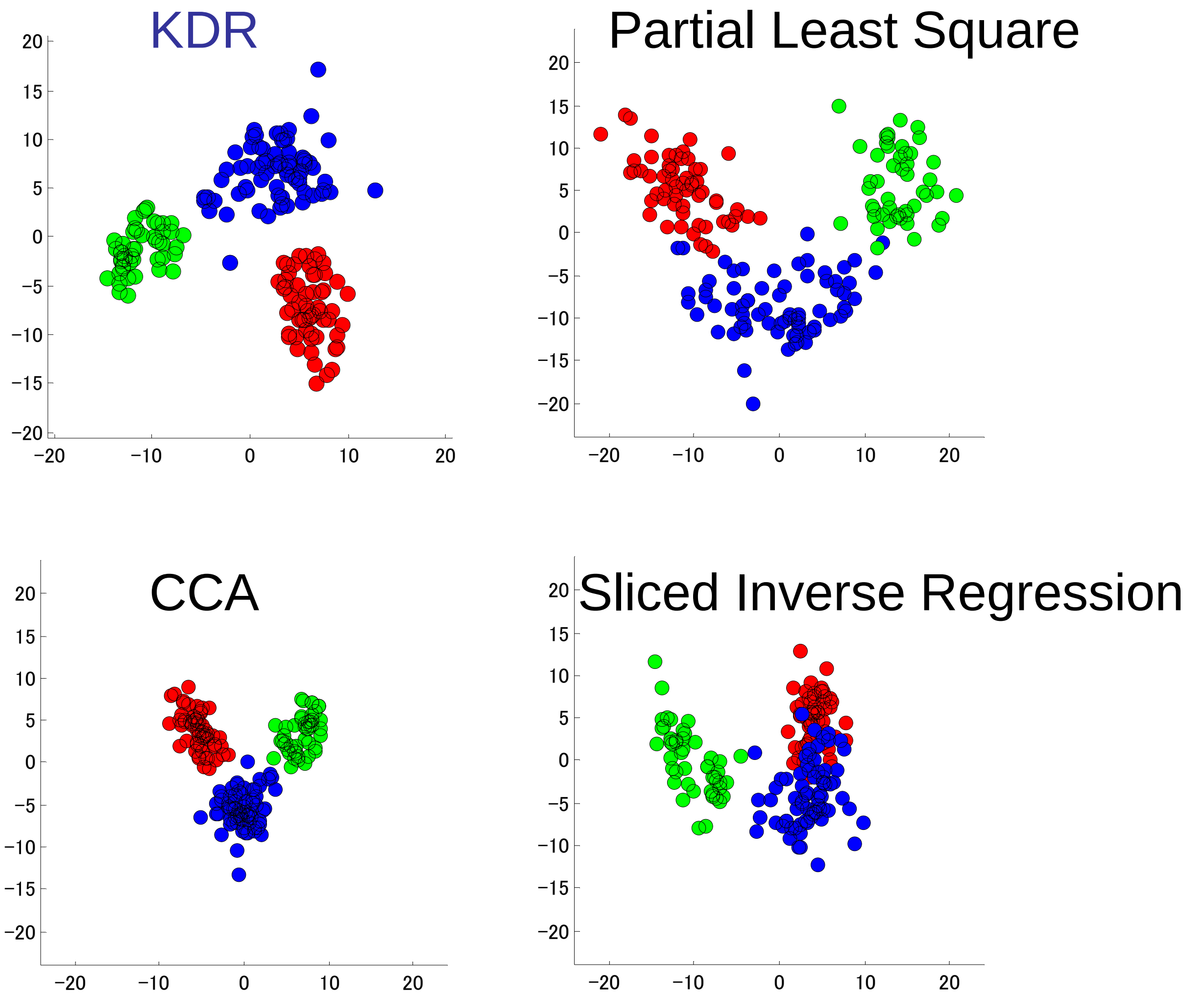
- Wine data

  Data

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

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

$$\sigma = 30$$



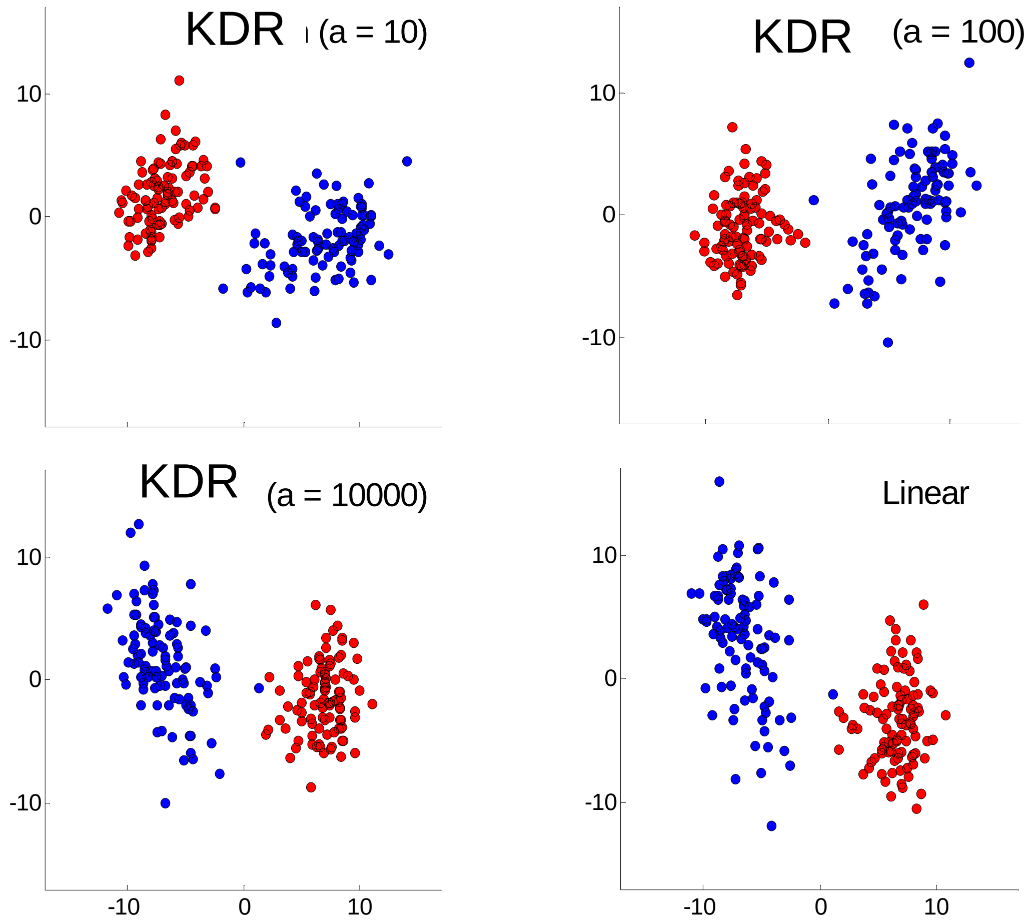
Swiss bank notes data

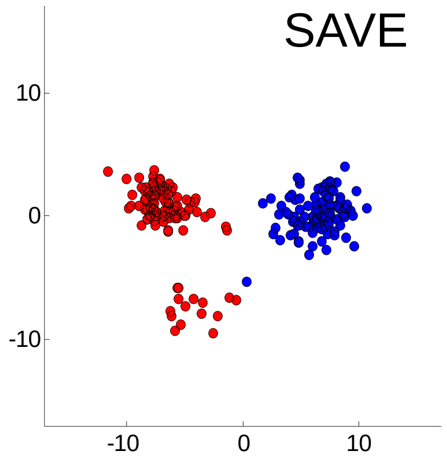
X: 6 dim. (measurements of each bank note)

Y: binary (genuine/counterfeit)

100 counterfeits ● and 100 genuine notes ●

$$k(z_1, z_2) = \exp(-\| z_1 - z_2 \|^2 / a)$$





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 kernel. Better than heuristics.
  - Choice of dimensionality for KDR.
  - Further asymptotic properties of the KDR estimator.

References

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

Fukumizu, K., F. Bach and M. Jordan. Kernel dimension reduction in regression. Tech. Report 715, Dept. Statistics, University of California, Berkeley, 2006.

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

Fukumizu, K., A. Gretton, X. Sun., and B. Schölkopf. Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems* 20:489-496. 2008.