Methods with Kernels

Statistical Data Analysis with Positive Definite Kernels

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

October 6-10, 2008, Kyushu University

Outline

Kernel Methodology

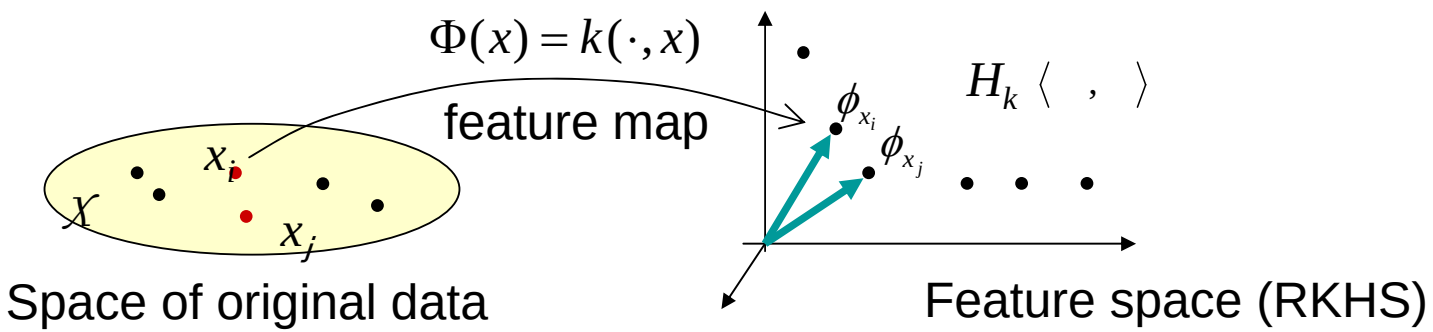
Kernel methodology: feature space by RKHS

Kernel methodology = Data analysis by transforming data into a high-dimensional feature space given by RKHS.

k: positive definite kernel.

$$\Phi : \mathcal{X} \to \mathcal{H}_k, \qquad x \mapsto \Phi(x) := k(\cdot, x)$$

$$\mathcal{X} \ni X_1, \ldots, X_N \quad \mapsto \quad \Phi(X_1), \ldots, \Phi(X_N) \in \mathcal{H}_k$$



Apply linear methods on RKHS – kernelization
The computation of the inner product is cheap.

Higher-order statistics by positive definite kernel

- A nonlinear kernel can include higher-order statistics.

  Example: Polynomial kernel on $\mathbb{R}$: $k(y, x) = (yx + 1)^d$.

  - Data are transformed as $k(\cdot, X_1), \ldots, k(\cdot, X_N) \in \mathcal{H}_k$.
  - Regarding $k(\cdot, X) = k(y, X)$ as a function of y,

    $$k(y, X) = X^d y^d + a_{d-1} X^{d-1} y^{d-1} + \cdots + a_1 X y + a_0 \qquad (a_i \neq 0).$$

  - $\{1, y, y^2, \ldots, y^d\}$ is a basis of $\mathcal{H}_k$.
  - With respect to this basis, the component of the feature vector $k(\cdot, X)$ is

    $$(X^d, a_{d-1} X^{d-1}, \ldots, a_1 X, a_0)^T.$$

    This includes the statistics $(X, X^2, \ldots, X^d)$.

- Similar nonlinear statistics appear in other kernels such as Gaussian, Lapacian, etc.

Kernel PCA I

- $X_1, \ldots, X_N$: data on $\mathcal{X}$.
- $k : \mathcal{X} \times \mathcal{X}$ positive definite kernel, $\mathcal{H}_k$: RKHS.
- Transform the data into $\mathcal{H}_k$ by $\Phi(x) = k(\cdot, x)$:

$$X_1, \ldots, X_N \quad \mapsto \Phi(X_1), \ldots, \Phi(X_N).$$

Kernel PCA ([SSM98]): Apply PCA on $\mathcal{H}_k$:

- Maximize the variance of the projection onto the unit vector f.

$$\max_{\|f\|=1} \mathrm{Var}[\langle f, \Phi(X)\rangle] = \max_{\|f\|=1} \frac{1}{N}\sum_{i=1}^{N}\left(\langle f, \Phi(X_i)\rangle - \frac{1}{N}\sum_{j=1}^{N}\langle f, \Phi(X_j)\rangle\right)^2$$

- It suffices to use $f = \sum_{i=1}^{n} a_i \tilde{\Phi}(X_i)$, where

$$\tilde{\Phi}(X_i) = \Phi(X_i) - \frac{1}{N}\sum_{j=1}^{N}\Phi(X_j).$$

The direction orthogonal to $\{\tilde{\Phi}(X_1), \ldots, \tilde{\Phi}(X_N)\}$ does not contribute.

Kernel PCA II

- The PCA solution:

$$\max a^T \tilde{K}^2 a \qquad \text{subject to} \quad a^T \tilde{K} a = 1,$$

where $\tilde{K}$ is $N \times N$ matrix with $\tilde{K}_{ij} = \langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle$.

$$\tilde{K} = k(X_i, X_j) - \tfrac{1}{N}\sum_{b=1}^{N} k(X_i, X_b) - \tfrac{1}{N}\sum_{a=1}^{N} k(X_a, X_j) + \tfrac{1}{N^2}\sum_{a,b=1}^{N} k(X_a, X_b).$$

$\tilde{K}$ is called a centered Gram matrix.

Note:

$$\tfrac{1}{N}\sum_{i=1}^{N} \langle f, \tilde{\Phi}(X_i) \rangle^2 = \tfrac{1}{N}\sum_{i=1}^{N} \langle \sum_{j=1}^{N} a_j \tilde{\Phi}(X_j), \tilde{\Phi}(X_i) \rangle^2 = \tfrac{1}{N} a^T \tilde{K}^2 a,$$

$$\|f\|^2 = \langle \sum_{i=1}^{n} a_i \tilde{\Phi}(X_i), \sum_{i=1}^{n} a_i \tilde{\Phi}(X_i) \rangle = a^T \tilde{K} a.$$

Kernel PCA III

- The p-th principal direction $f^{(p)} = \sum_{i=1}^{N} \alpha_i^{(p)} \tilde{\Phi}(X_i)$ is given by

$$\max \alpha^{(p)T} \tilde{K}^2 \alpha^{(p)} \quad \text{subj. to} \quad \begin{cases} \alpha^{(p)} \tilde{K} \alpha^{(p)} = 1 \\ \alpha^{(p)} \tilde{K} \alpha^{(a)} = 0 \quad (a = 1, \ldots, p-1). \end{cases}$$

Principal components of kernel PCA

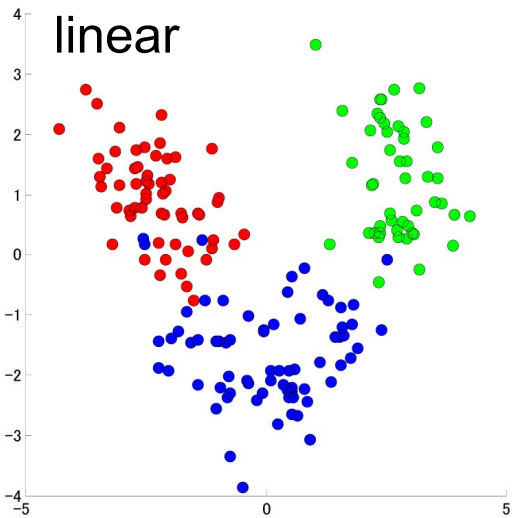
Let $\tilde{K} = \sum_{p=1}^{N} \lambda_p u^{(p)} u^{(p)T}$ is the eigen decomposition $(\lambda_1 \geq \cdots \geq \lambda_N \geq 0)$.
The p-th principal component of the data X_i is

$$\langle \tilde{\Phi}(X_i), \sum_{j=1}^{N} \alpha_j^{(p)} \tilde{\Phi}(X_j) \rangle = \sum_{j=1}^{N} \sqrt{\lambda_p} u_i^{(p)},$$

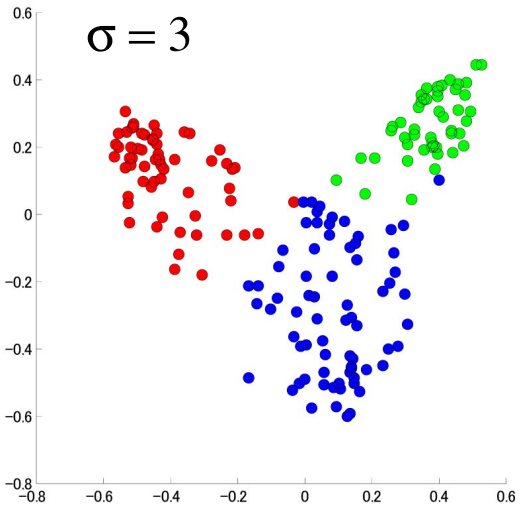
Kernel PCA: numerical examples

- Wine data (from UCI repository [MA94]).
- 178 data of 13 dimension. They represents chemical measurements of different wine.
- There are three classes, which correspond to types of wine.
- The classes are shown in different colors, but not used for the analysis.

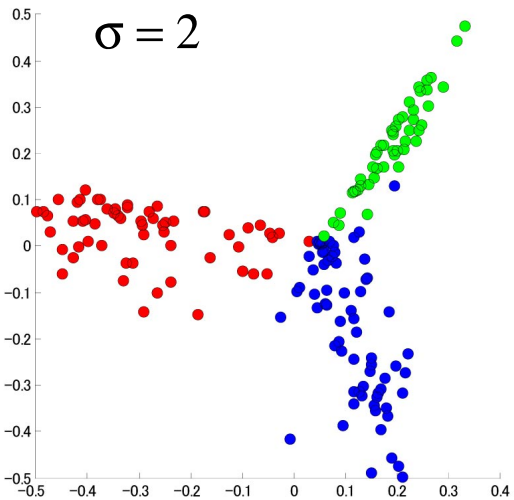
Linear PCA

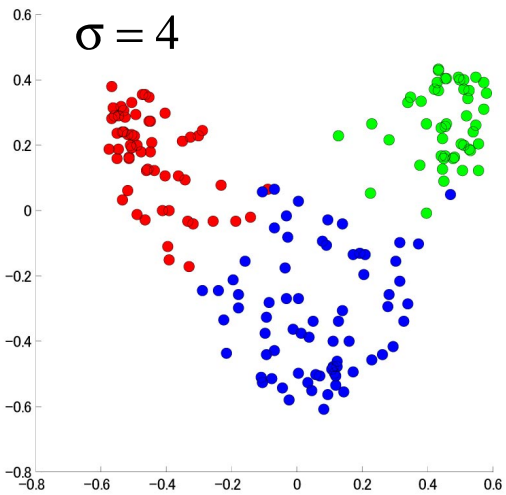


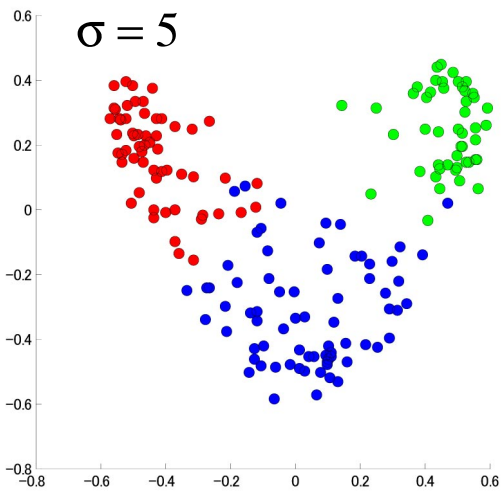
KPCA Gaussian kernel



KPCA with Gaussian kernels. $k(x, y) = \exp\{-\frac{1}{\sigma^2}\|x - y\|^2\}$.





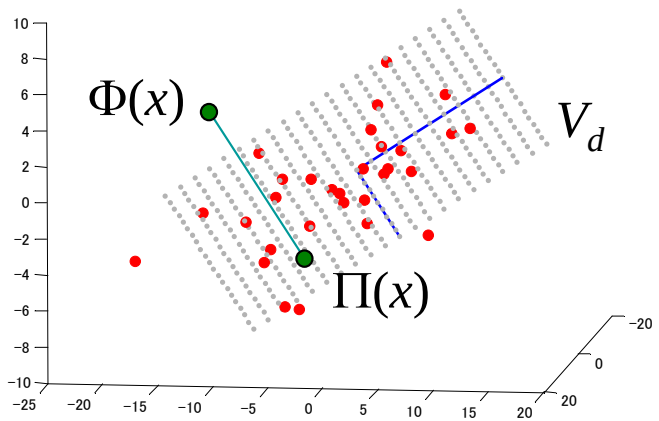


Application of KPCA to noise reduction I

- $X_1, \ldots, X_N$: data, $\mapsto \Phi(X_1), \ldots, \Phi(X_N)$: data in RKHS.
- V_d: principal subspace of $\mathcal{H}_k$ spanned by $f^{(1)}, \ldots, f^{(d)}$.
- $\Pi(x)$ $(\in \mathcal{H}_k)$: orthogonal projection of $\Phi(x)$ onto V_d.
- Find a point y in the original space such that

$$y = \arg\min_{y \in \mathcal{X}} \|\Phi(y) - \Pi(x)\|_{\mathcal{H}_k}.$$

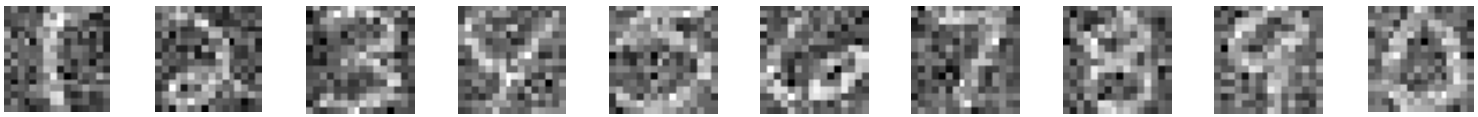
Note: $\Pi(x)$ is not necessarily in the image of embedding Φ.



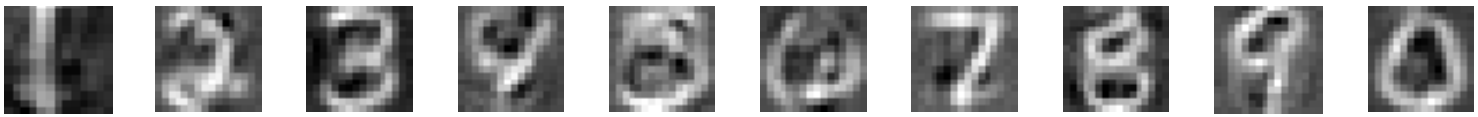
Application of KPCA to noise reduction II

USPS hand-written digits data:
7191 images of hand-written digits of 16×16 pixels.

Sample of original images (not used for experiments)



Sample of noisy images



Sample of denoised images (linear PCA)

Sample of denoised images (kernel PCA, Gaussian kernel)

Generated by Matlab Stprtool (by V. Franc).

Properties of kernel PCA

- Nonlinear features can be considered.
- The results depend on the choice of kernel and kernel parameters. Interpreting the results may not be straightforward.
- Can be used for a preprocessing of other analysis like classification. (Dimension reduction / feature extraction)
- How to choose a kernel and kernel parameter?
  - Cross-validation may be possible, in general.
  - If it is a preprocessing, the performance of the final analysis should be maximized.

Canonical correlation analysis I

Canonical correlation analysis (CCA)

- Linear dependence of two multivariate.
  - Data $(X_1, Y_1), \ldots, (X_N, Y_N)$
  - X_i: m-dimensional, Y_i: ℓ-dimensional.
- Find the directions a and b so that the correlation between the projections of X onto a and that of Y onto b is maximized:

$$\rho = \max_{a \in \mathbb{R}^m, b \in \mathbb{R}^\ell} \frac{\mathrm{Cov}[a^T X, b^T Y]}{\sqrt{\mathrm{Var}[a^T X]\mathrm{Var}[b^T Y]}} = \max_{a \in \mathbb{R}^m, b \in \mathbb{R}^\ell} \frac{a^T \widehat{V}_{XY} b}{\sqrt{a^T \widehat{V}_{XX} a}\sqrt{b^T \widehat{V}_{YY} b}},$$

where $\widehat{V}_{XX}$, $\widehat{V}_{YY}$, and $\widehat{V}_{XY}$ are the sample variance (covariance) matrices.

Canonical correlation analysis II

- Optimization:

$$\max a^T \widehat{V}_{XY} b \quad \text{subject to } a^T \widehat{V}_{XX} a = b^T \widehat{V}_{YY} b = 1.$$

- Solution is obtained by the largest ρ for the generalized eigenproblem:

$$\begin{pmatrix} O & \widehat{V}_{XY} \\ \widehat{V}_{YX} & O \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \rho \begin{pmatrix} \widehat{V}_{XX} & O \\ O & \widehat{V}_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

Derivation: Lagrange multiplier

$$\max a^T \widehat{V}_{XY} b + \frac{\mu}{2}(a^T \widehat{V}_{XX} a - 1) + \frac{\nu}{2}(b^T \widehat{V}_{YY} b - 1).$$

From $\partial/\partial a = \partial/\partial b = 0$,

$$V_{XY} b + \mu V_{XX} a = 0, \qquad V_{YX} a + \nu V_{YY} b = 0.$$

$\mu = \nu$ is derived. Set $\rho = -\mu = -\nu$.

Kernel CCA I

Kernel CCA: kernelization of CCA ([Aka01, MRB01, BJ02]).

- Data: $(X_1, Y_1), \ldots, (X_N, Y_N)$.
  - X_i, Y_i: arbitrary variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ (resp.).
- Embedding: prepare kernels $k_{\mathcal{X}}$ on $\mathcal{X}$ and $k_{\mathcal{Y}}$ on $\mathcal{Y}$.
  $X_1, \ldots, X_N \quad \mapsto \quad \Phi_{\mathcal{X}}(X_1), \ldots, \Phi_{\mathcal{X}}(X_N) \in \mathcal{H}_{k_{\mathcal{X}}}$.
  $Y_1, \ldots, Y_N \quad \mapsto \quad \Phi_{\mathcal{Y}}(Y_1), \ldots, \Phi_{\mathcal{Y}}(Y_N) \in \mathcal{H}_{k_{\mathcal{Y}}}$.
- Apply CCA on $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$.

$$\max_{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\sum_{i=1}^N \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}}{\sqrt{\sum_{i=1}^N \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}}^2} \sqrt{\sum_{i=1}^N \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}^2}}$$

where

$$\tilde{\Phi}_{\mathcal{X}}(X_i) = \Phi_{\mathcal{X}}(X_i) - \frac{1}{N} \sum_{j=1}^N \Phi_{\mathcal{X}}(X_j), \quad \text{and } \tilde{\Phi}_{\mathcal{Y}}(Y_i) \text{ similar.}$$

Kernel CCA II

- We can assume $f = \sum_{i=1}^{N} \alpha_i \tilde{\Phi}_{\mathcal{X}}(X_i)$ and $g = \sum_{i=1}^{N} \beta_i \tilde{\Phi}_{\mathcal{Y}}(Y_i)$.

$$\rho = \max_{\alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^N} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \tilde{K}_X^2 \alpha} \sqrt{\beta^T \tilde{K}_Y^2 \beta}},$$

  $\tilde{K}_X$ and $\tilde{K}_Y$ are the centered Gram matrices.

- Regularization:
  Canonical correlation in N dimensional space with N data is ill-posed with correlation 1.

$$\max_{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\sum_{i=1}^{N} \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}}{\sqrt{\sum_{i=1}^{N} \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}}^2 + \varepsilon_N \|f\|^2} \sqrt{\sum_{i=1}^{N} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}^2 + \varepsilon_N \|g\|^2}}$$

Kernel CCA III

- Kernel CCA

$$\begin{pmatrix} O & \tilde{K}_X\tilde{K}_Y \\ \tilde{K}_Y\tilde{K}_X & O \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} \tilde{K}_X^2 + \varepsilon_N K_X & O \\ O & \tilde{K}_Y^2 + \varepsilon_N K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

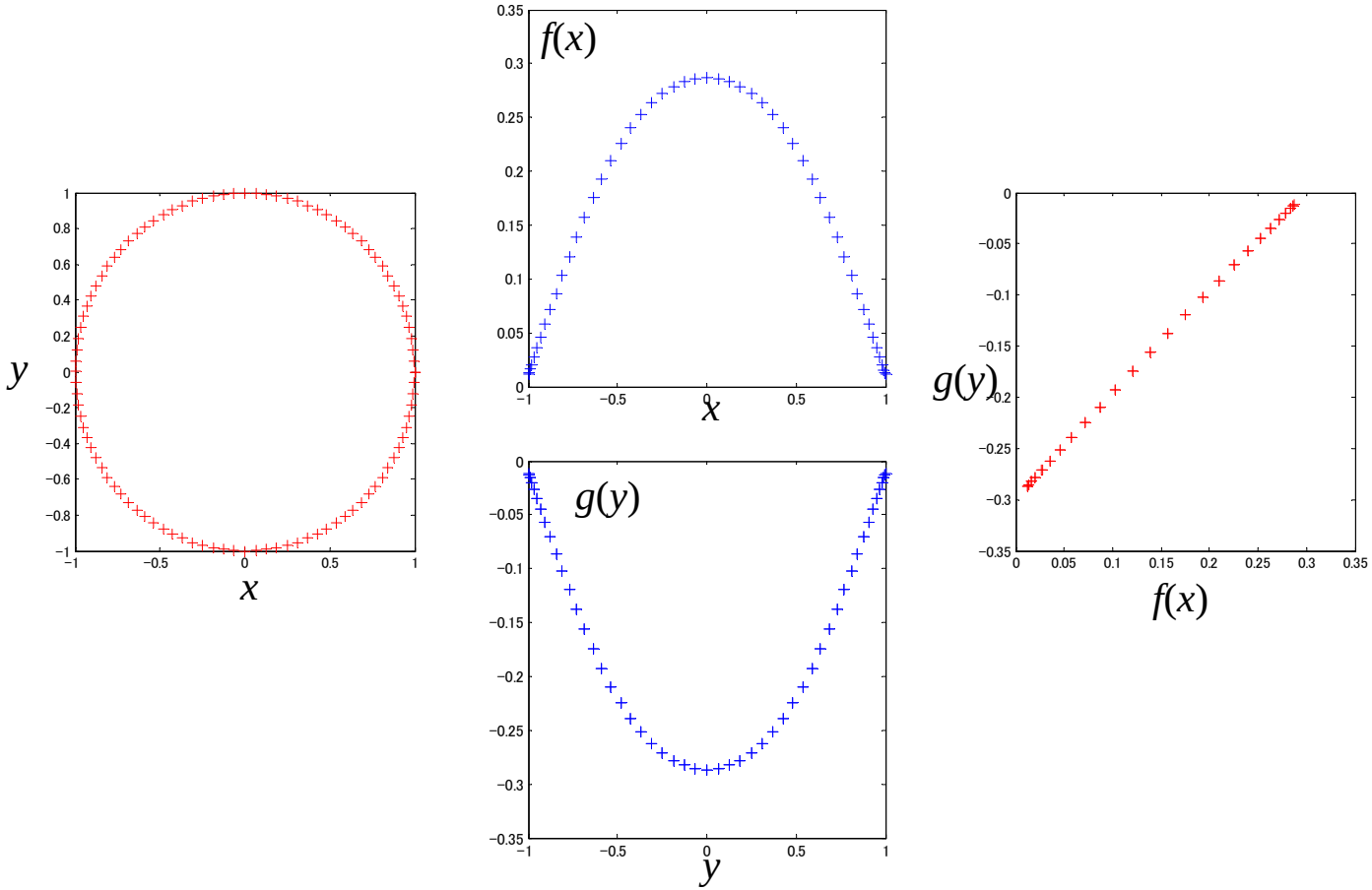
  The Solution is obtained as a generalized eigenproblem.

- The multiple feature vectors (second, third, eigenvectors) can be also obtained.

- Remark:
  - The results of kernel CCA depends on the kernels and ε_N.
  - The consistency is known if ε_N decreases sufficiently slowly as $N \to \infty$ [FBG07].

Toy example of Kernel CCA

X, Y: one-dimensional. Gaussian RBF kernels are used.



Application of Kernel CCA

Application of kernel CCA to image retrieval ([HSST04]).

Idea: use d eigenvectors $f_1, \ldots, f_d$ and $g_1, \ldots, g_d$ as the feature spaces which contain the dependence between X and Y.

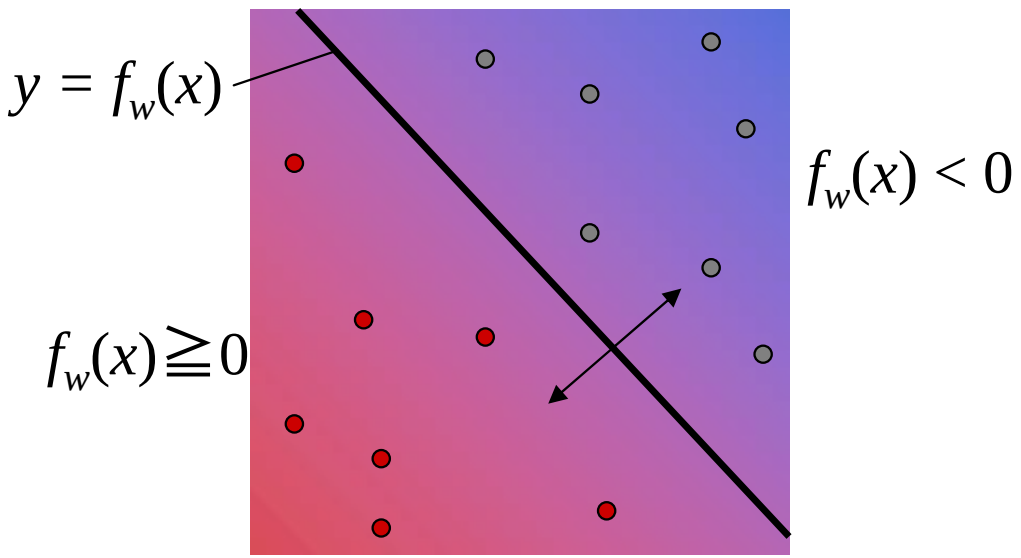
- X_i: image, Y_i: text (extracted from webpages).
- Compute the d-eigenvectors $f_1, \ldots, f_d$ and $g_1, \ldots, g_d$ by kernel CCA.
- Compute the feature vectors by projections $\xi_i = (\langle \Phi_{\mathcal{X}}(X_i), f_a \rangle_{\mathcal{H}_{\mathcal{X}}})_{a=1}^d \in \mathbb{R}^d$ for all images.
- For a new text Y_{new}, compute the feature $\zeta = (\langle \Phi_{\mathcal{Y}}(Y_{new}), g_a \rangle_{\mathcal{H}_{\mathcal{Y}}})_{a=1}^d \in \mathbb{R}^d$, and output the image

$$\arg\max_i = \xi_i^T \zeta.$$

Linear classifier

- $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: explanatory variable (m-dimensional)
  - $Y_i \in \{+1, -1\}$ binary,
- Linear classifier

$$f(x) = sgn\left(w^T x + b\right)$$



Large margin classifier I

Linear support vector machine (in $\mathbb{R}^m$)

- Assumption: the data is linearly separable.
- Large margin criterion:
  Among infinite number of separating hyperplanes, choose the one to give the largest margin.
  - Margin = distance of two classes measured along the direction of w.
  - The classifying hyperplane is the middle of the margin.

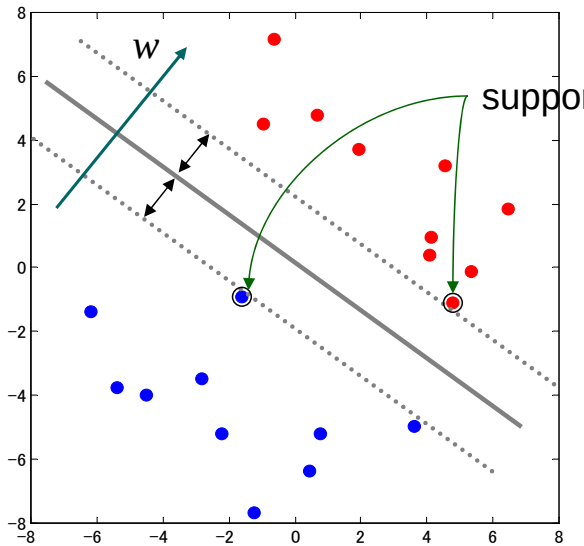
Large margin classifier II

To fix a scale, assume

$$\begin{cases} \min(w^T X_i + b) = 1 & i : Y_i = +1 \\ \max(w^T X_i + b) = -1 & i : Y_i = -1 \end{cases}$$

Then,

$$\text{Margin} = \frac{2}{\|w\|}$$



Large margin classifier III

- Large margin linear classifier

$$\max \frac{1}{\|w\|} \qquad \text{subj. to} \begin{cases} w^T X_i + b \geq 1 & \text{if } Y_i = +1, \\ w^T X_i + b \leq -1 & \text{if } Y_i = -1. \end{cases}$$

Equivalently,

Linear support vector machine (hard margin)

$$\min_{w,b} \|w\|^2 \qquad \text{subject to} \qquad Y_i(w^T X_i + b) \geq 1 \quad (\forall i).$$

- Quadratic objective function with linear constraints $\Longrightarrow$ free from local minima!
- This optimization can be numerically solved with the standard quadratic programming (QP, quadratic objective function with lienar constraints. Discussed later). Software packages are available.

SVM with soft margin

Relax the separability assumption. The linear separability is too restrictive in practice.

- Hard constraint: $Y_i(w^T X_i + b) \geq 1$
- Soft constraint: $Y_i(w^T X_i + b) \geq 1 - \xi_i \quad (\xi_i \geq 0)$

Linear support vector machine (soft margin)

$$\min_{w,b,\xi_i} \|w\|^2 + C\sum_{i=1}^N \xi_i \qquad \text{subj. to} \qquad \begin{cases} Y_i(w^T X_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

- The optimization is still QP.
- C is a hyper-parameter, which we have to decide.

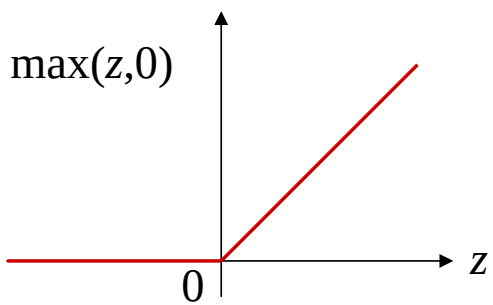
Soft margin as regularization

- Soft margin linear SVM is equivalent to the following regularization problem ($\lambda = 1/C$):

$$\min_{w,b} \sum_{i=1}^{N} \big(1 - Y_i(w^T X_i + b)\big)_+ + \lambda \|w\|^2$$

where

$$(z)_+ = \max(z, 0)$$



- $\ell(f(x), y) = (1 - yf(x))_+$: hinge loss.

Tikhonov Regularization

General theory of regularization

- When the solution of the optimization

$$\min_{\alpha \in A} \Omega(\alpha)$$

  ($A \subset \mathcal{H}$) is not unique or stable, a regularization technique is often used.
- Tikhonov regularization: add a regularization term (or penalty term), e.g.,

$$\min_{\alpha \in A} \ \Omega(\alpha) + \lambda \|\alpha\|^2.$$

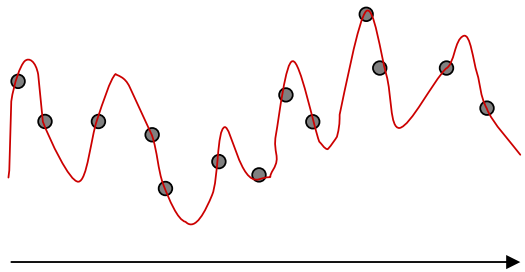
  $\lambda > 0$: regularization coefficient.
- The solution is often unique and stable.
- Other regularization terms, such as $\|\alpha\|$, are also possible, but differentiability may be lost.

Tikhonov Regularization II

- Example
  - Ill-posed problem:

$$\min_f (Y_i - f(X_i))^2.$$

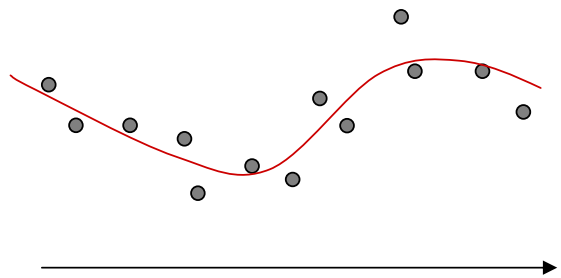
Many f give zero error, if f is taken from a large space.



  - Regularized objective function

$$\min_f (Y_i - f(X_i))^2 + \lambda \|f\|^2$$

finds a unique solution, which is often smoother.



SVM with kernels I

Kernelization of linear SVM

- $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: arbitrary covariate taking values in $\mathcal{X}$,
  - $Y_i \in \{+1, -1\}$ binary,
- k: positive definite kernel on $\mathcal{X}$. $\mathcal{H}$: associated RKHS.
- $\Phi(X_i) = k(\cdot, X_i)$: transformed data in $\mathcal{H}$.
- Large margin linear classifier on RKHS

$$f(x) = sgn\big(\langle h, \Phi(x)\rangle_{\mathcal{H}} + b\big) = sgn(h(x) + b).$$

Objective function (soft margin):

$$\min_{h,b,\xi_i} \|h\|^2_{\mathcal{H}} + C\sum_{i=1}^N \xi_i \quad \text{subj. to} \quad \begin{cases} Y_i(\langle h, \Phi(X_i)\rangle + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \end{cases}$$

or equivalently

$$\min_{h,b} \sum_{i=1}^N \big(1 - Y_i(\langle h, \Phi(X_i)\rangle + b)\big)_+ + \lambda\|h\|^2$$

SVM with kernels II

- It suffices to assume

$$h = \sum_{i=1}^{N} c_i \Phi(X_i)$$

The orthogonal direction only increases the regularization term without changing the first term of

$$\min_{h,b} \sum_{i=1}^{N} \left(1 - Y_i(\langle h, \Phi(X_i)\rangle + b)\right)_+ + \lambda \|h\|^2.$$

- In this case,

$$\|h\|^2 = \sum_{i,j=1}^{N} c_i c_j k(X_i, X_j),$$

$$\langle h, \Phi(X_i)\rangle = \sum_{j=1}^{N} c_j k(X_i, X_j).$$

SVM with kernels III

In summary,

SVM with kernel

$$\min_{c_i,b,\xi_i} \sum_{i,j=1}^{N} c_i c_j k(X_i, X_j) + C\sum_{i=1}^{N} \xi_i,$$

$$\text{subj. to} \quad \begin{cases} Y_i(\sum_{j=1}^{N} k(X_i, X_j)c_j + b) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

- The optimization is numerically solved with QP.
- The dual form is simpler to solve (discussed later.)
- The parameter C and the kernel are often chosen by cross-validation.

Demonstration of SVM

Webpages for SVM Java applet

- `http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml`
- `http://www.eee.metu.edu.tr/~alatan/Courses/Demo/AppletSVM.html`

Results on character recognition

MNIST: Handwritten digit recognition
28×28 binary pixels.

60000 training data
10000 test data

	k-NN Euclid	10PCA + quad.	RBF + lin.	LeNet-4	LeNet-5	SVM poly4	RS-SVM poly5
Test error (%)	5.0	3.3	3.6	1.1	0.95	1.1	1.0

Taken from [LBBH98]

Mini-summary on SVM

- Kernel trick (a common property of kernel methods):
  - linear classifier on RKHS.
  - The computation of inner product is easy.
- Large margin criterion
  - May not be the Bayes optimal, but causes other good properties.
- Quadratic programming:
  - The objective function is solved by the standard quadratic programming.
- Sparse representation:
  - The classifier is represented by a small number of support vectors.
- Regularization:
  - The soft margin objective function is equivalent to the margin loss with regularization.

Representer theorem I

Minimization problems on RKHS

$$\min_{f \in \mathcal{H}_k} (Y_i - f(X_i))^2 + \lambda \|f\|^2 \qquad \text{(ridge regression)},$$

$$\min_{f \in \mathcal{H}_k, b} \sum_{i=1}^{N} \big(1 - (Y_i f(X_i) + b)\big)_+ + \lambda \|f\|^2 \qquad \text{(SVM)}.$$

We have seen that the solution can be taken from

$$f = \sum_{i=1}^{N} \alpha_i k(\cdot, X_i).$$

Representer theorem II

- General problem:
  - $\mathcal{H}_k$: RKHS with associated with a positive definite kernel k.
  - $X_1, \ldots, X_N$, $Y_1, \ldots, Y_N$: data.
  - $h_1(x), \ldots, h_m(x)$: fixed functions.
  - $\Psi : [0\ \infty) \to \mathbb{R}$: non-decreasing function (regularization term).

  Minimization

$$\min_{f \in \mathcal{H}, c \in \mathbb{R}^m} L\left(\{X_i\}_{i=1}^N, \{Y_i\}_{i=1}^N, \{f(X_i) + \sum_{a=1}^m c_a h_a(X_i)\}_{i=1}^N\right) + \Psi(\|f\|).$$

Representer theorem

The solution of the above minimization is achieved by a function of the form

$$f = \sum_{i=1}^N \alpha_i k(\cdot, X_i).$$

- The optimization in an high (or infinite) dimensional space can be reduced to the optimization in a subspace of N dimension (sample size).

Proof of the representer theorem

- Decomposition:

$$\mathcal{H}_k = H_0 \oplus H_0^{\perp},$$

$H_0 = \text{span}\{k(\cdot, X_1), \ldots, k(\cdot, X_N)\}$, $H_0^{\perp}$: orthogonal complement. Decompose

$$f = f_0 + f^{\perp}$$

accordingly.

- Because

$$\langle f^{\perp}, k(\cdot, X_i) \rangle = 0,$$

the loss function L does not change by replacing f with f_0.

- The second term:

$$\|f_0\| \leq \|f\| \qquad \Longrightarrow \qquad \Psi(\|f_0\|) \leq \Psi(\|f\|).$$

- Thus, the optimum f can be in the space H_0.

Other kernel methods

- Kernel PLS
- Support vector regression (SVR)
- Kernel logistic regression
- Kernel FDA (Fisher discriminant analysis)
- Kernel K-means clustering
- Other variants of SVM (ν-SVM, one-class SVM etc.). etc...

Structured data

Positive definite kernels can be defined on an arbitrary set.

Kernel methods can be applied to any type of data (vector / non-vectorial).

Structured data: non-vectorial data with some structure such as strings, tees, graphs, and so on.

Special kernels are studied for each domain.

Choice of Kernel

How to choose a kernel?

- Reflect knowledge on the problem as much as possible. (structured data)
- For supervised learning such as SVM, use cross-validation.
- For unsupervised learning such as kernel PCA and kernel CCA, there are no theoretically guaranteed methods.

  Suggestions: make a relevant supervised method and use cross-validation.

Summary of Section 3

- Various classical linear methods of data analysis can be kernelized – linear algorithms on RKHS.

  Kernel PCA, SVM, kernel CCA, kernel FDA, etc.

- The solution often has the form

$$f = \sum_{i=1}^{N} \alpha_i k(\cdot, X_i)$$

  (representer theorem).

- The problem is reduced to operations on Gram matrices of the sample size N.

- The kernel methods can be applied to any type of data including non-vectorial (structured) data, such as graphs, strings, etc, if a positive definite kernel is provided.

References I

[Aka01] Shotaro Akaho.
A kernel method for canonical correlation analysis.
In *Proceedings of International Meeting on Psychometric Society (IMPS2001)*, 2001.

[BJ02] Francis R. Bach and Michael I. Jordan.
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48, 2002.

[FBG07] Kenji Fukumizu, Francis R. Bach, and Arthur Gretton.
Statistical consistency of kernel canonical correlation analysis.
Journal of Machine Learning Research, 8:361–383, 2007.

[HSST04] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor.
Canonical correlation analysis: An overview with application to learning methods.
Neural Computation, 16:2639–2664, 2004.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner.
Gradient based learning applied to document recognition.
Proceedings of IEEE, 86(11):2278–2324, 1998.

References II

[MA94] Patrick M. Murphy and David W. Aha.

UCI repository of machine learning databases.

Technical report, University of California, Irvine, Department of Information and Computer Science. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1994.

[MRB01] Thomas Melzer, Michael Reiter, and Horst Bischof.

Nonlinear feature extraction using generalized canonical correlation analysis.

In *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 353–360, 2001.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller.

Nonlinear component analysis as a kernel eigenvalue problem.

Neural Computation, 10:1299–1319, 1998.