

Given a set of centered observations i.e. $\sum_{k=1}^{M} x_k = 0$ Diagonalization of covariance matrix C $C = \frac{1}{M} \sum_{j=1}^{M} x_j x_j^T$ (1) via the solution of eigenvalue equation $\lambda v = Cv$ where $\lambda \ge 0$ and $v \in \mathbb{R}^N \setminus \{0\}$ (2) therefore $Cv = \frac{1}{M} \sum_{j=1}^{M} x_j x_j^T v$

Linear PCA

since $(xx^T)v = (x \cdot v)x$ derivation (?)

A dot product formulation

$$Cv = \frac{1}{M} \sum_{j=1}^{M} (x_j \cdot v) x_j = \lambda v$$
 $(x_j \cdot v)$ - scalar

Therefore all solutions v lie in the span of $(x_1 \dots x_M)$ And eigenvalue equation for each data point:

$$\lambda(x_k \cdot v) = (x_k \cdot Cv) \qquad \forall k = 1, \dots, M$$

Non-Linear PCA

 $\ensuremath{\textit{F}}\xspace$ – feature space, related to the input space by a non-linear map Φ

 $\Phi: R^{\scriptscriptstyle N} \to F$

Given a set of centered observations i.e. $\sum_{k=1}^{M} \Phi(x_k) = 0$

Covariance matrix C in F

$$\overline{C} = \frac{1}{M} \sum_{j=1}^{M} \Phi(x_j) \Phi(x_j)^T$$

Again we have to solve the eigenvalue equation

$$\lambda V = \overline{C}V$$
 where $\lambda \ge 0$ and $V \in F \setminus \{0\}$

By the analogy with the linear PCA the solutions *V* lie in the span of non-linear input mappings

$$(\Phi(x_1)\ldots\Phi(x_M))$$

Non-Linear PCAEigenvalue equation for each data point: $\lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \overline{C}V)$ $\forall k = 1, \dots, M$ Since V are linearly related to inputs $\Phi(x_k)$ we can define coefficients α $V = \sum_{i=1}^{M} \alpha_i \Phi(x_i)$ (4)

Combining (3), (4) and defining matrix *K* (macro step !):

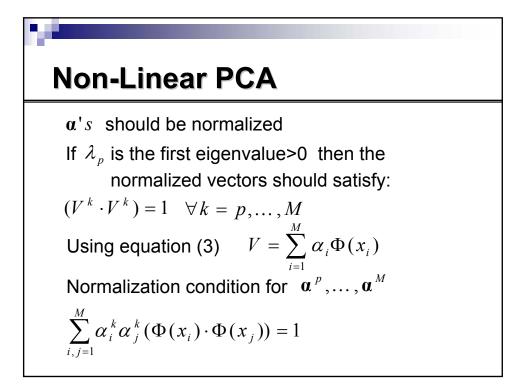
$$K_{ii} = (\Phi(x_i) \cdot \Phi(x_i))$$

We get:

$$M\lambda a = Ka$$

(4)

K is a positive semidefinite \rightarrow diagonalize it to get the solutions for the equation (4)



Non-Linear PCA

$$\sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k (\Phi(x_i) \cdot \Phi(x_j)) = 1 \qquad k = p, \dots, M$$

$$\sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k K_{ij} = 1$$

$$(\mathbf{a}^k \cdot K \mathbf{a}^k) = 1$$

$$\lambda_k (\mathbf{a}^k \cdot \mathbf{a}^k) = 1 \quad \text{the normalization condition}$$

Extracting non-linear principal components Let x be a test point, $\Phi(x)$ is image of x in F

$$(V^k \cdot \Phi(x)) = \sum_{i=1}^M \alpha_i^k (\Phi(x_i) \cdot \Phi(x))$$

Three steps of non-linear PCA:

1. Compute the dot product matrix K

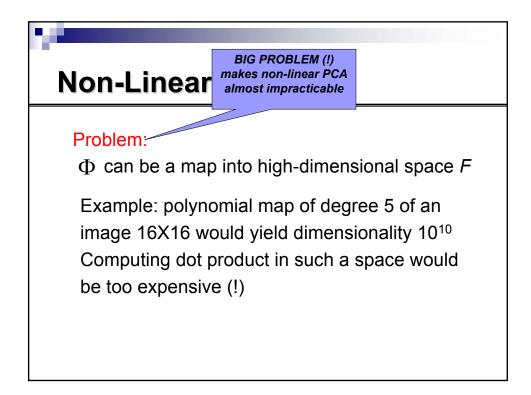
$$K_{ij} = (\Phi(x_i) \cdot \Phi(x_j))$$

2. Compute Eigenvectors of *K* and normalize them in *F*

$$\lambda_k(\boldsymbol{\alpha}^k\cdot\boldsymbol{\alpha}^k)=1$$

3. Compute projections of a test point onto the Eigenvectors

$$PC_{k}(x) = (V^{k} \cdot \Phi(x)) = \sum_{i=1}^{M} \alpha_{i}^{k} (\Phi(x_{i}) \cdot \Phi(x))$$



Kernel PCA

Solution:

kernel PCA where dot products can be represented using the kernel function

$$k(x, y) = (\Phi(x) \cdot \Phi(y))$$

This allows to compute $(\Phi(x) \cdot \Phi(y))$ without explicitly mapping x into *F*

Kernel function

How does a kernel work:

Suppose $\Phi(x)$ is a quadratic basis function and is the input vector of dimensionality d then the full quadratic expansion is...

	1	├ constant term			
	$\sqrt{2}x_1$				
	$\sqrt{2} x_2$	linear terms			
	$\sqrt{2} x_d$				
$\Phi(x) =$	$x \frac{2}{2}$	pure quadratic terms			
	x_d^2				
	$\begin{array}{c} \sqrt{2} x_{1} x_{2} \\ \sqrt{2} x_{1} x_{3} \end{array}$				
	$\begin{vmatrix} \vdots \\ \sqrt{2}x \\ x \end{vmatrix}$				
	$ \begin{array}{c} \sqrt{2} x_{1} x_{d} \\ \sqrt{2} x_{2} x_{3} \\ \cdot \end{array} $	quadratic cross-terms			
	$\left \begin{array}{c} \frac{1}{\sqrt{2}} x_{2} x_{d} \right $				
	$\begin{bmatrix} \vdots \\ \sqrt{2} x_{d-1} x_{d} \end{bmatrix}$	Number of terms =(d+2)(d+1)/2			

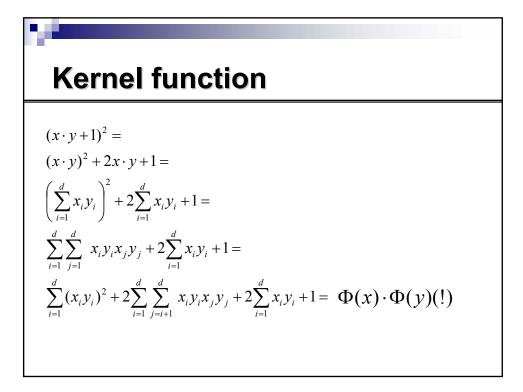
$$\Phi(x) \bullet \Phi(y) = \begin{cases} 1 & 1 & 1 & 1 \\ \sqrt{2} x_{1} & \sqrt{2} y_{2} & 1 & 1 \\ \sqrt{2} y_{2} & 1 & \sqrt{2} y_{2} & 1 \\ \sqrt{2} y_{2} & 1 & \sqrt{2} y_{2} & 1 \\ \sqrt{2} y_{2} & 1 & \sqrt{2} y_{2} & 1 \\ \sqrt{2} y_{1} & y_{1}^{2} & y_{$$

Kernel function

$$\Phi(x) \cdot \Phi(y) = 1 + \sum_{i=1}^{d} 2x_i y_i + \sum_{i=1}^{d} x_i^2 y_i^2 + \sum_{i=1}^{d} \sum_{j=i+1}^{d} 2x_i x_j y_i y_j$$

Let's consider another function of x and y:

$$(x \cdot y + 1)^2$$



Kernel function

Definition of a kernel function:

 $k(x, y) = (\Phi(x) \cdot \Phi(y))$

So $(x \cdot y + 1)^2$ is the kernel function of x and y given $\Phi()$ is the mapping function into quadratic feature space *F*

Polynomial kernel function of degree p:

 $(x \cdot y + 1)^p$

Kernel function								
	-	nomial kerr nd number						
		1						
Polynomial	# terms $\Phi(x)$	# terms d=100	Kernel	# terms kernel	# terms d=100			
Quadratic	d²/2	5000	$(x \cdot y + 1)^2$	d/2	50			
Cubic	d ³ /6	166,000	$(x \cdot y + 1)^3$	d/2	50			
Quartic	d ⁴ /24	~4,000,000	$(x \cdot y + 1)^4$	d/2	50			

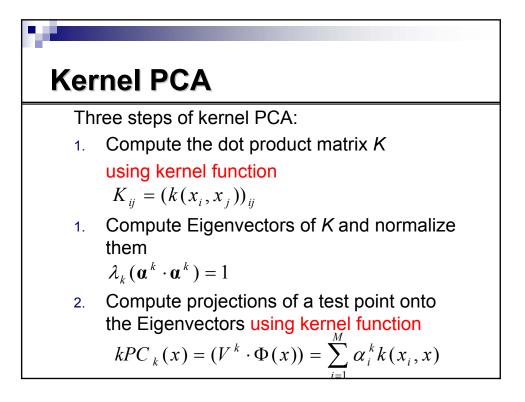
Other kernel functions

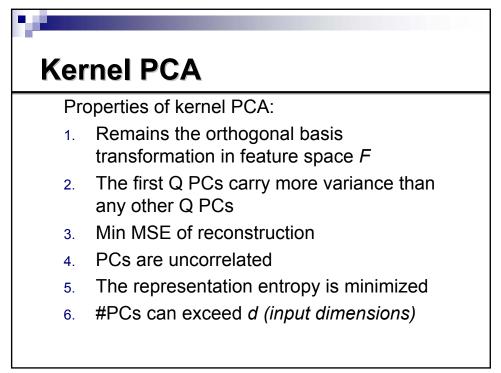
Radial basis

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\delta^2}\right)$$

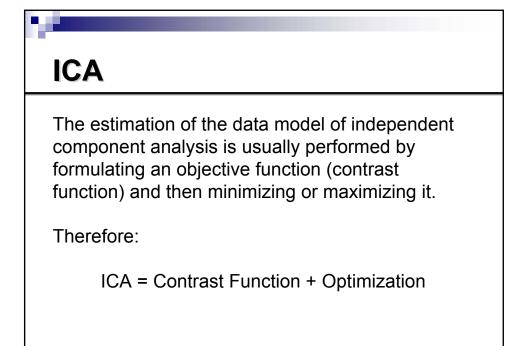
Neural Network type

$$k(x, y) = \tanh((x \cdot y) + b)$$





ICA Independent component analysis (ICA) decomposes the multivariate data $y \in R^N$ into a linear sum of statistically independent components: $y = \sum_{i=1}^{N} x_i a_i = Ax$ where x_i is the basis coefficient (source) and a_i is the basis vector The task is to estimate parameters *A* from data



F-correlation

In the paper by Bach and Jordan (2001) "Kernel ICA", a new contrast function, based on *F*-correlation, was developed

This new function is based on the non-linear function space not on just one function

F-correlation

F-correlation – measures dependence between x_1 and x_2 using correlation of functions of the variables $f_1(x_1)$ and $f_2(x_2)$ for f_1 and f_2 belonging to some space *F*

$$\rho_F = \max_{f_1, f_2 \in F} corr(f_1(x_1), f_2(x_2))$$
$$\rho_F = \max_{f_1, f_2 \in F} \frac{cov(f_1(x_1), f_2(x_2))}{(var f_1(x_1))^{1/2} (var f_2(x_2))^{1/2}}$$

F-correlation

If $\rho_F = 0$ and *F* is large enough then x_1 and x_2 are independent Large enough?

If F contains the Fourier basis i.e. all functions of the form:

$$x \mapsto e^{i\omega x}$$

where $\omega \in R$ How to make this tractable?

"kernelize" *F*-correlation

"kernelized" F-correlation is equivalent to canonical correlation

Canonical Correlation

Given two multivariate random variables $\ x_1 \in R^{N_1}$ and $\ x_2 \in R^{N_2}$ CCA finds the pair of directions $\,\mathcal{W}_{2}^{}$ and $\,\mathcal{W}_{1}^{}^{}$ with maximum correlation.

$$\rho(x_1, x_2) = \max_{w_1, w_2} corr(w_1^T x_1, w_2^T x_2)$$
$$= \max_{w_1, w_2} \frac{w_1^T C_{12} w_2}{(w_1^T C_{11} w_1)^{\frac{1}{2}} (w_2^T C_{22} w_2)^{\frac{1}{2}}}$$

Where $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$ is the covariance matrix of (x₁, x₂)

Canonical Correlation

The CCA reduces to the generalized eigenvalue problem:

$$\begin{pmatrix} 0 & C_{12} \\ C_{21} & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \rho \begin{pmatrix} C_{11} & 0 \\ 0 & C_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

This problem has N₁+N₂ generalized eigenvalues ρ "Kernelized" F-correlation is equivalent to canonical correlation

RKHS

RKHS – reproducing kernel Hilbert spaces Let K(x, y) be a Mercer kernel on $X = R^p$ i.e. a function for which the Gram matrix

$$K_{ij} = K(x_i, x_j)$$

is positive definite for any collection $\,\{x_i\}_{i=1,\ldots,N}\,$ in $\,X$

Corresponding to any such kernel K there is a map Φ from X to a feature space F such that

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

RKHS is the instantiation of *F* with the following property called "reproducing property" $f(x) = \langle K(\cdot, x), f \rangle$ $\forall f \in F$ If $\Phi(x) = K(\cdot, x)$ is a map from the input space into the RKHS then $\langle \Phi(x), \Phi(y) \rangle = \langle K(\cdot, x), K(\cdot, y) \rangle = K(x, y)$

Kernel

Isotropic Gaussian kernel – Mercer kernel with the feature space F the space of smooth functions

$$K(x, y) = \exp(-\frac{1}{2\sigma^2} ||x - y||^2)$$

Theorem 1

Theorem 1

Let x_1 and x_2 be random variables in $X = R^p$ Let K_1 and K_2 be Mercer kernels with feature maps Φ_1 and Φ_2 and feature spaces $F_1, F_2 \in R^X$ Then the canonical correlation between $\Phi_1(x_1)$ and $\Phi_2(x_2)$ which is defined as

$$\rho_F = \max_{(f_1, f_2) \in F_1 \times F_2} \operatorname{corr}(\langle \Phi_1(x_1), f_1 \rangle, \langle \Phi_2(x_2), f_2 \rangle)$$

is equal

$$\rho_F = \max_{(f_1, f_2) \in F_1 \times F_2} corr(f_1(x_1), f_2(x_2))$$

Theorem 2

Theorem 2 (Independence and F-correlation) If F is the RKHS corresponding to Gaussian kernel $\rho_F = 0$ iff x₁ and x₂ are independent

$\begin{aligned} \mathcal{F}(x_1, x_2) &= \max_{w_1, w_2 \in \mathbb{R}^N} \frac{w_1^T K_1 K_2 w_2}{(w_1^T K_1^2 w_1)^{\frac{1}{2}} (w_1^T K_2^2 w_1)^{\frac{1}{2}}} \\ \text{Where } K_1 \text{ and } K_2 \text{ are Gram matrices of } x_1 \text{ and } x_2 \\ \text{This is equivalent to performing CCA on two vectors with covariance matrix} \\ \begin{pmatrix} K_1^2 & K_1 K_2 \\ K_2 K_1 & K_2^2 \end{pmatrix} \end{aligned}$

Kernelized CCA

The kernelized CCA reduces to the generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

