



# Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group

Data61 | CSIRO

and

College of Engineering and Computer Science

The Australian National University

Canberra

February – June 2017

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

## Outlines

Overview

Introduction

Linear Algebra

Probability

Linear Regression 1

Linear Regression 2

Linear Classification 1

Linear Classification 2

Kernel Methods

Sparse Kernel Methods

Neural Networks 1

Neural Networks 2

Continuous Latent Variables

Autoencoders

Graphical Models 1

Graphical Models 2

Graphical Models 3

Sampling

Mixture Models and EM 1

Mixture Models and EM 2

Sequential Data 1

Sequential Data 2

Combining Models



## Part XIII

# *Principal Component Analysis*

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



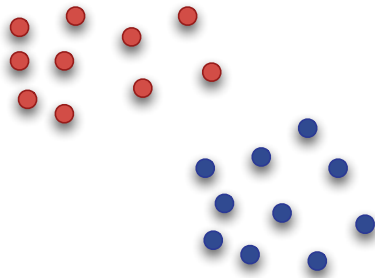
- Deep architectures get stuck in local minima or plateaus
- As architecture gets deeper, more difficult to obtain good generalisation
- Hard to initialise random weights well
- 1 or 2 hidden layers seem to perform better
- 2006: Unsupervised pre-training

## Motivation

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



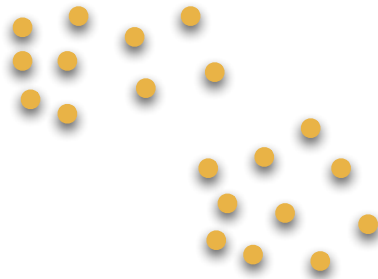
- Given are pairs of data  $x_i \in \mathcal{X}$  and targets  $t_i \in \mathcal{T}$  in the form  $(x_i, t_i)$ , where  $i = 1..N$ .
- Learn a mapping between the data  $X$  and the target  $t$  which generalises well to new data.

## Motivation

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



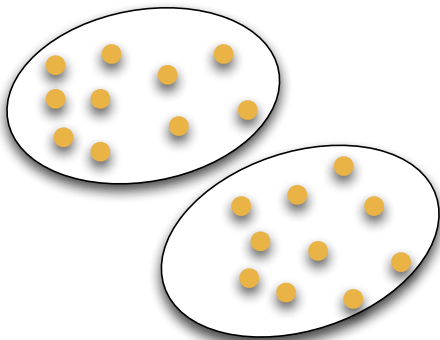
- Given only the data  $x_i \in \mathcal{X}$ .
- Discover (=learn) some interesting structure inherent in the data  $X$ .

## Motivation

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



- Given only the data  $x_i \in \mathcal{X}$ .
- Discover (=learn) some interesting structure inherent in the data.

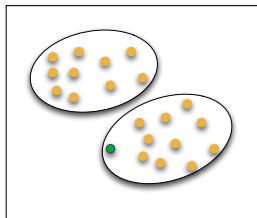
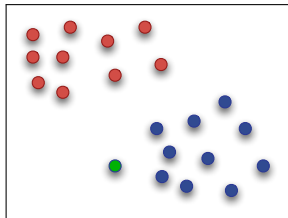
## Motivation

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*

# Testing - Supervised versus Unsupervised Learning



## Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis

# Recall: Fisher's Linear Discriminant



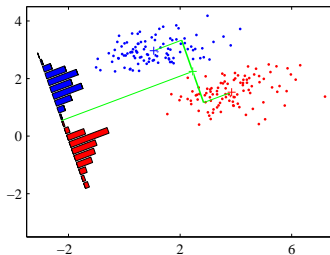
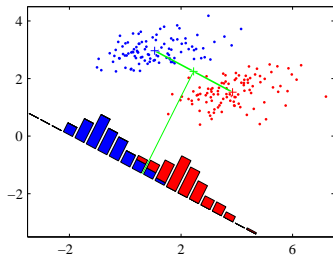
Samples from two classes in a two-dimensional input space and their histogram when projected to two different one-dimensional spaces.

## Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis







- Main goal : **dimensionality reduction**
- Many applications in visualisation, feature extraction, signal processing, data compression . . .
- Example: Use hand-written digits (binary data) and place them into a larger frame ( $100 \times 100$ ) varying the position and the rotation angle.
- Data space size = 10 000.
- But data live on a three-dimensional manifold ( $x$ ,  $y$ , and the rotation angle).



*Motivation*

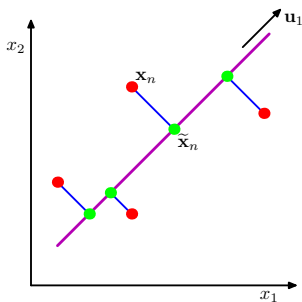
*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*

# Principal Component Analysis (PCA)

- Idea: Linearly project the data points onto a lower dimensional subspace such that
  - the variance of the projected data is maximised, or
  - the distortion error from the projection is minimised.
- Both formulations lead to the same result.
- Need to find the lower dimensional subspace, called the **principal subspace**.



Motivation

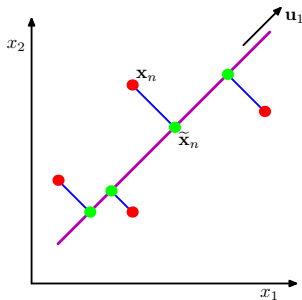
Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis

# Principal Component Analysis (PCA)

- Given  $N$  observations  $\mathbf{x}_n \in \mathbb{R}^D$ ,  $n = 1, \dots, N$ .
- Project onto a space with dimensionality  $M < D$  while maximising the variance.
- More advanced : How to calculate  $M$  from the data. Therefore here:  $M$  is fixed.
- Consider a 1-dimensional subspace spanned by some unit vector  $\mathbf{u}_1 \in \mathbb{R}^D$ ,  $\mathbf{u}_1^T \mathbf{u}_1 = 1$ .





- Each data point  $\mathbf{x}_n$  is then projected onto a scalar value  $\mathbf{u}_1^T \mathbf{x}_n$ .
- The mean of the projected data is  $\mathbf{u}_1^T \bar{\mathbf{x}}$  where  $\bar{\mathbf{x}}$  is the sample mean

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n.$$

- The variance of the projected data is then

$$\frac{1}{N} \sum_{n=1}^N \{ \mathbf{u}_1^T \mathbf{x}_n - \mathbf{u}_1^T \bar{\mathbf{x}} \}^2 = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$$

with the covariance matrix

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T.$$

Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis



- Maximising  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$  under the constraint  $\mathbf{u}_1^T \mathbf{u}_1 = 1$  (why do we need to bound  $\mathbf{u}_1$  ?) leads to the Lagrange equation

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



- Maximising  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$  under the constraint  $\mathbf{u}_1^T \mathbf{u}_1 = 1$  (why do we need to bound  $\mathbf{u}_1$  ?) leads to the Lagrange equation

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

which has a stationary point

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



- Maximising  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$  under the constraint  $\mathbf{u}_1^T \mathbf{u}_1 = 1$  (why do we need to bound  $\mathbf{u}_1$  ?) leads to the Lagrange equation

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

which has a stationary point if  $\mathbf{u}_1$  is an eigenvector of  $\mathbf{S}$  with eigenvalue  $\lambda_1$ .

$$\mathbf{S} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1.$$

- The variance is then  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda_1$ .
- Variance is maximised if  $\mathbf{u}_1$  is the eigenvector of the covariance  $\mathbf{S}$  with the largest eigenvalue.

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



- Continue maximising the variance amongst all possible directions orthogonal to those already considered.
- The optimal linear projection onto a  $M$ -dimensional space for which the variance is maximised is defined by the  $M$  eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_M$  of the covariance matrix  $\mathbf{S}$  corresponding to the  $M$  largest eigenvalues  $\lambda_1, \dots, \lambda_M$ .
- Is this subspace always uniquely defined?
  
- Not if  $\lambda_M = \lambda_{M+1}$ .

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*





- The distortion between data points  $\mathbf{x}_n$  and their projection  $\tilde{\mathbf{x}}_n$

$$J = \frac{1}{N} \sum_{n=1}^N \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$$

is minimised if the variance is maximised.

- The distortion error is then

$$J = \sum_{i=M+1}^D \lambda_i$$

where  $\lambda_i, i = M + 1, \dots, D$  are the **smallest** eigenvalues of the covariance matrix  $\mathbf{S}$ .

- In signal processing we speak of the **signal space** (principal subspace) and the **noise space** (orthogonal to the principal subspace).

Motivation

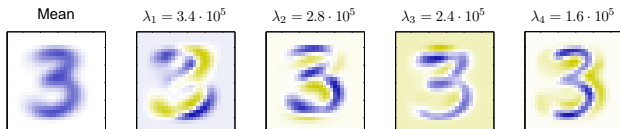
Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis



- The eigenvectors of the covariance matrix are elements of the original vector space  $u_i \in \mathbb{R}^D$ .
- If the input data are images, the eigenvectors are also images.



The mean and the first four eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_4$  of a set of handwritten digits of 'three'.

Blue corresponds to positive values, white is zero and yellow corresponds to negative values.

Motivation

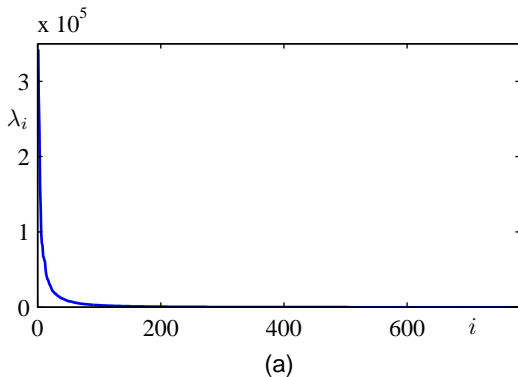
Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis



- The eigenvalues of the covariance matrix express the variance of the data set in the direction of the corresponding eigenvectors.



Plot of the eigenvalue spectrum for the digits of three data set.

Motivation

Principal Component  
Analysis

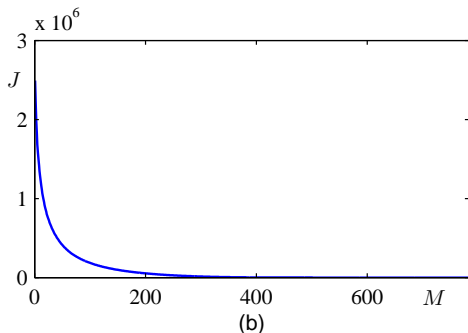
Principal Component  
Analysis (PCA)

Independent Component  
Analysis



- The sum of the eigenvalues of the covariance matrix of the discarded directions express the distortion error.

$$J = \frac{1}{N} \sum_{n=1}^N \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$$



Plot of the distortion error versus the number of dimension of the subspace considered for projection.

Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis



- The approximated data vector  $\tilde{\mathbf{x}}_n$  can be written in the form

$$\tilde{\mathbf{x}}_n = \bar{\mathbf{x}} + \sum_{i=1}^M (\mathbf{u}_i^T (\mathbf{x}_n - \bar{\mathbf{x}})) \mathbf{u}_i$$

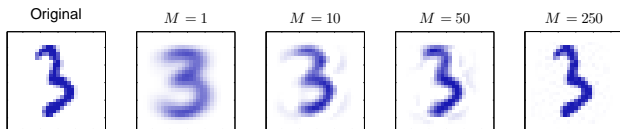
- Codebook :  $M + 1$  vectors of dimension  $D$  ( $\bar{\mathbf{x}}$  and  $\mathbf{u}_i$ ).
- Compressed  $\mathbf{x}_n$  :  $M$  factors  $\mathbf{u}_i^T (\mathbf{x}_n - \bar{\mathbf{x}})$

Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis



Reconstruction of an image retaining  $M$  principal components.



- Standardise certain features of a data set (for instance as a preprocessing step to subsequent algorithms expecting these features).
- Usually, individual standardisation: each variable (dimension) has zero mean and unit variance. But variables are still correlated.
- PCA can do more: create **decorrelated** data (covariance is the identity; also called **whitening** or **sphering** of the data)
- Write the eigenvector equation for the covariance matrix  $\mathbf{S}$

$$\mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{L}$$

where  $\mathbf{L}$  is the diagonal matrix of (positive!) eigenvalues.

- Transform the original data by

$$\mathbf{y}_n = \mathbf{L}^{-1/2} \mathbf{U}^T (\mathbf{x}_n - \bar{\mathbf{x}})$$

- The set  $\{\mathbf{y}_n\}$  has mean zero and covariance given by the identity.

Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis



- Transform the original data by

$$\mathbf{y}_n = \mathbf{L}^{-1/2} \mathbf{U}^T (\mathbf{x}_n - \bar{\mathbf{x}})$$

- Mean of the set  $\{\mathbf{y}_n\}$

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^N \mathbf{y}_n &= \frac{1}{N} \sum_{n=1}^N \mathbf{L}^{-1/2} \mathbf{U}^T (\mathbf{x}_n - \bar{\mathbf{x}}) \\ &= \mathbf{L}^{-1/2} \mathbf{U}^T \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}}) = \mathbf{0} \end{aligned}$$

- Covariance of the set  $\{\mathbf{y}_n\}$

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^N \mathbf{y}_n \mathbf{y}_n^T &= \frac{1}{N} \sum_{n=1}^N \mathbf{L}^{-1/2} \mathbf{U}^T (\mathbf{x}_n - \bar{\mathbf{x}}) (\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{U} \mathbf{L}^{-1/2} \\ &= \mathbf{L}^{-1/2} \mathbf{U}^T \mathbf{S} \mathbf{U} \mathbf{L}^{-1/2} \\ &= \mathbf{L}^{-1/2} \mathbf{U}^T \mathbf{U} \mathbf{L}^{-1/2} \\ &= \mathbf{I} \end{aligned}$$

Motivation

Principal Component  
Analysis

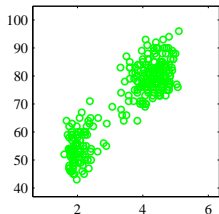
Principal Component  
Analysis (PCA)

Independent Component  
Analysis

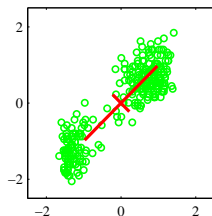
# PCA - The Effect of Whitening



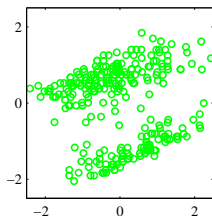
- Compare standardising and whitening of a data set.
- (b) also shows the principal axis of the normalised data set plotted as red lines over the range  $\pm\lambda_i^{1/2}$ .



Original data  
(note the different  
axis).



Standardising to  
zero mean and unit  
variance.



Whitening to  
achieve unit  
covariance.

Motivation

Principal Component  
Analysis

Principal Component  
Analysis (PCA)

Independent Component  
Analysis





- **Kernel PCA**

- Use  $\Phi(x)$  as features, and express in terms of kernel matrix  $\mathbf{K}$
- The covariance matrix  $\mathbf{S}$  and the (centered) kernel matrix  $\mathbf{K}$  has the same eigenvalues.

- **Probabilistic PCA**

- Explicitly model latent variable  $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ .
- Mean value of observed variable is given by  $\mathbf{W}\mathbf{z} + \mu$
- Conditional distribution of observed variable

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*



# Independent Component Analysis - Overview

- Assume we have  $K$  signals and  $K$  recordings, each recording containing a mixture of the signals.
- ‘Cocktail party’ problem :  $K$  people speak at the same time in a room, and  $K$  microphones pickup a mixture of what they say.
- Given unknown source signals  $S \in \mathbb{R}^{N \times K}$  and an unknown **mixing matrix**  $\mathbf{A}$ , producing the observed data  $X \in \mathbb{R}^{N \times K}$

$$X = SA$$

- Can we recover the original signals (Blind Source Separation)?
- Yes, under the assumption that
  - at most one of the signals is Gaussian distributed.
  - we don’t care for the amplitude (including the sign).
  - we don’t care for the order of the recovered signals.
  - we have at least as many observed mixtures as signals, the matrix  $\mathbf{A}$  has full rank and can be inverted.



# Independence versus Uncorrelatedness

- Independence

$$p(x_1, x_2) = p(x_1)p(x_2)$$

- Uncorrelatedness (defined via a zero covariance)

$$\mathbb{E}[x_1 x_2] - \mathbb{E}[x_1] \mathbb{E}[x_2] = 0$$

- Independence implies Uncorrelatedness (prove it!).
- BUT Uncorrelatedness does NOT imply Independence.
- Example: Draw the pair  $(x_1, x_2)$  with equal probability from the set  $\{(0, 1), (0, -1), (1, 0), (-1, 0)\}$ .
- Then  $x_1$  and  $x_2$  are uncorrelated because  $\mathbb{E}[x_1] = \mathbb{E}[x_2] = \mathbb{E}[x_1 x_2] = 0$ .
- But  $x_1$  and  $x_2$  are NOT independent

$$p(x_1 = 0, x_2 = -1) = \frac{1}{4}$$

$$p(x_1 = 0)p(x_2 = -1) = \frac{1}{2} \times \frac{1}{4}$$

# Independent Component Analysis - Overview



- Uncorrelated variables are not necessarily independent.
- ICA maximises the **statistical independence** of the estimated components.
- Find  $A^{-1}$  in such a way that

$$S = XA^{-1}$$

the columns of  $S$  are maximally independent.

- Several definitions for statistical independence possible.
- Based on the concept of what is 'nongaussian'.
- Central Limit Theorem: The distribution of a sum of independent random variables tends toward a Gaussian distribution (under certain conditions).
- FastICA algorithm.

*Motivation*

*Principal Component  
Analysis*

*Principal Component  
Analysis (PCA)*

*Independent Component  
Analysis*