T3A: Machine Learning Algorithms

Master of Science in AI and Master of Science in Data Science

@ UPSaclay

2024/2025.

Faïcel Chamroukhi

chamroukhi.com

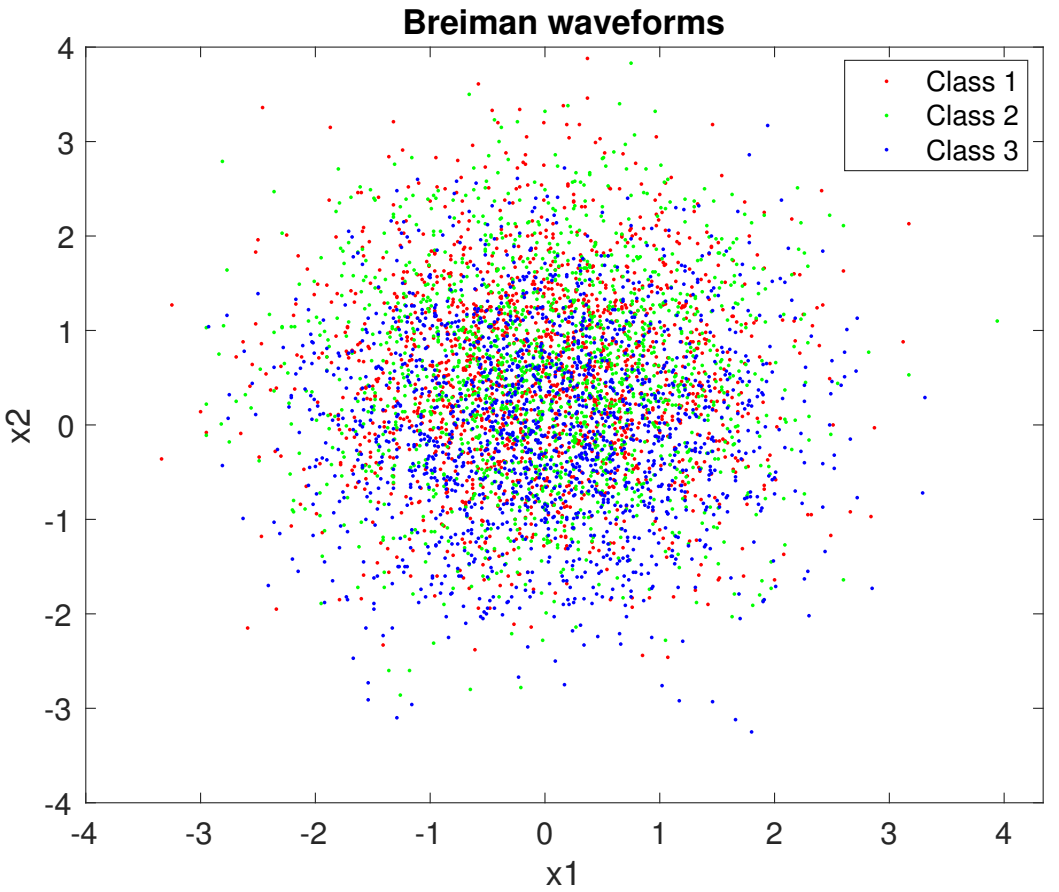
Outline

1 Introduction

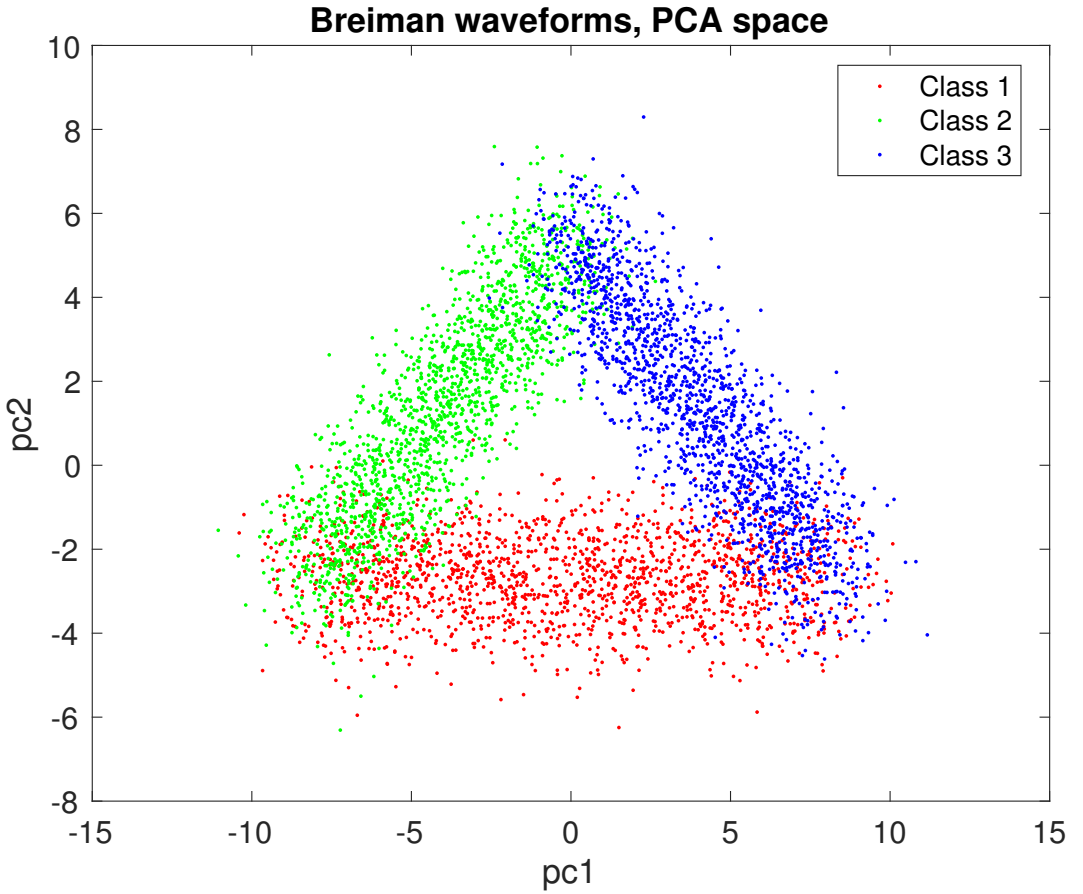
2 Dimensionality Reduction

3 Latent data models for dimensionality reduction

Representation



Representation



Unsupervised Learning

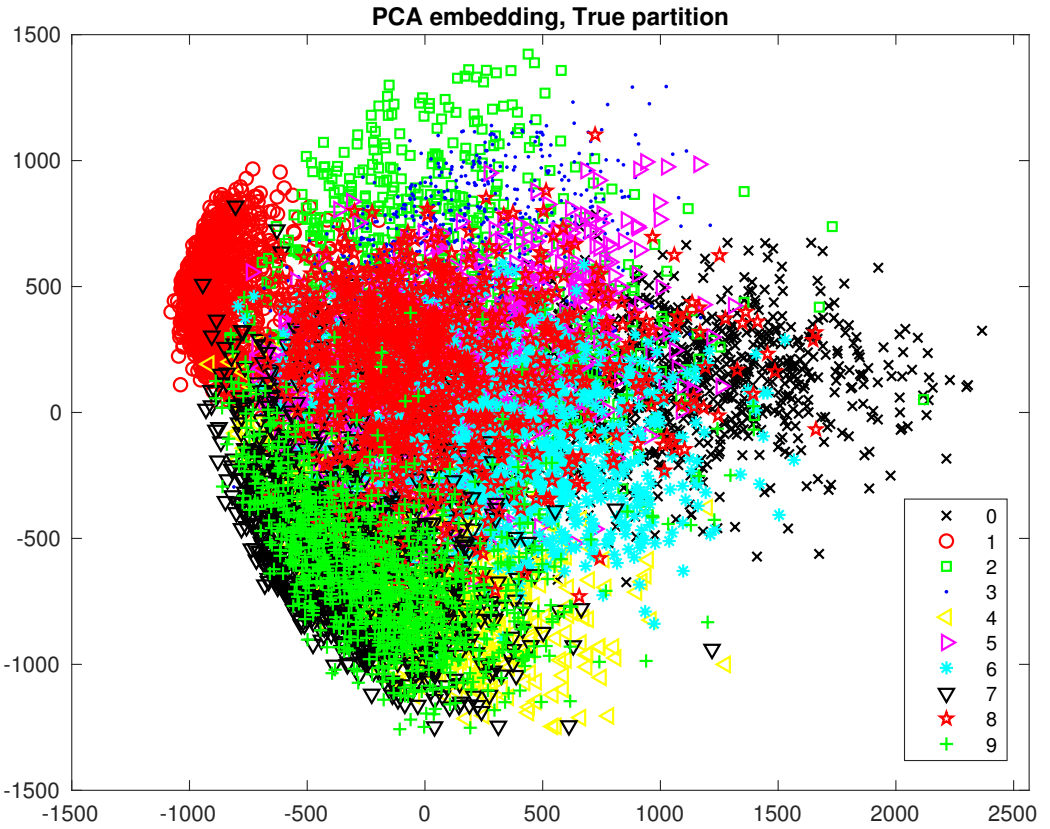
université PARIS-SACLAY

Clustering / Representation / Data viz / Dimensionality reduction

Unsupervised Analysis

université PARIS-SACLAY

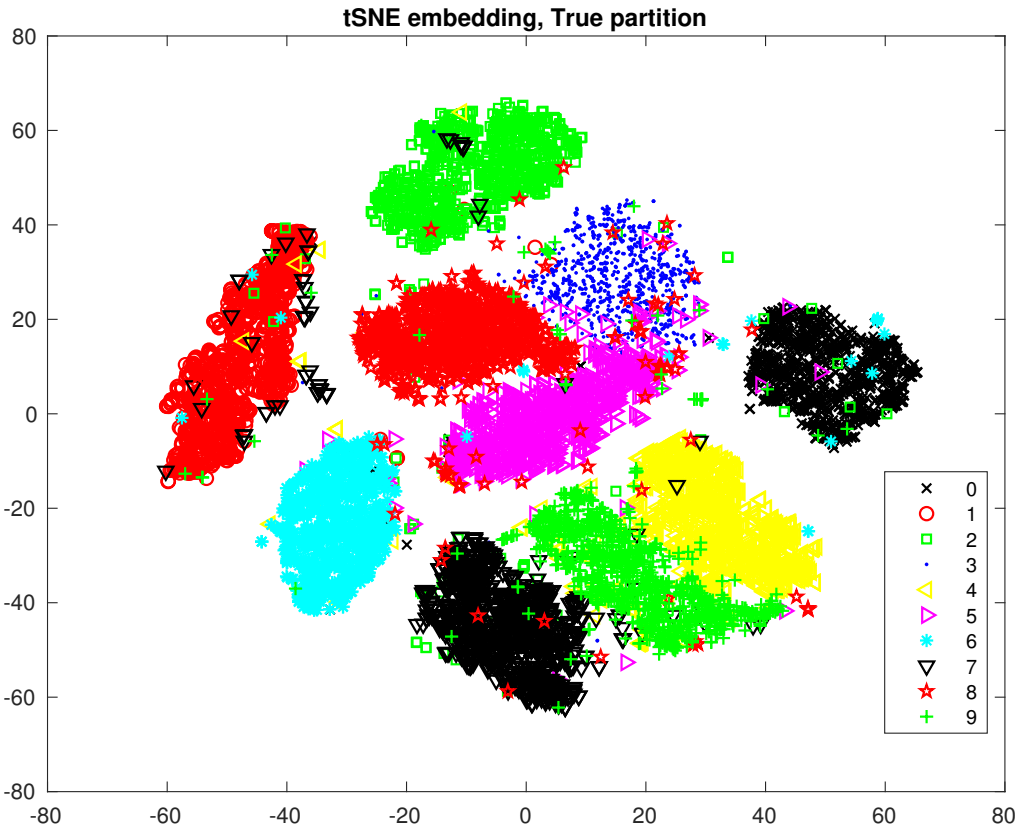
Representation / Data viz / Dimensionality reduction



Unsupervised Analysis

université PARIS-SACLAY

Representation / Data viz / Dimensionality reduction



Dimensionality reduction

université PARIS-SACLAY

- Unsupervised learning for dimensionality reduction
- Principal Component Analysis (PCA)
- Probabilistic PCA (PPCA)
- Factor Analysis (FA)

reduction

université PARIS-SACLAY

- Dimensionality reduction for high dimensional data (for representation/visualization etc)
- Principal Component Analysis (PCA) [Pearson, 1901, Hotelling, 1933],
- Probabilistic PCA [Tipping and Bishop, 1997, 1999, Roweis, 1998]
- Factor Analysis (FA)[Spearman, 1904, Thurstone, 1947],

Principal Component Analysis (PCA)

université PARIS-SACLAY

- PCA is a linear projection which maximizes the variance in the projected space [Hotelling, 1933].

Consider a sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_i \in \mathbb{R}^d$.
$\Rightarrow$ The aim is to project the data onto a space having dimensionality $M < d$ while maximizing the variance of the projected data.

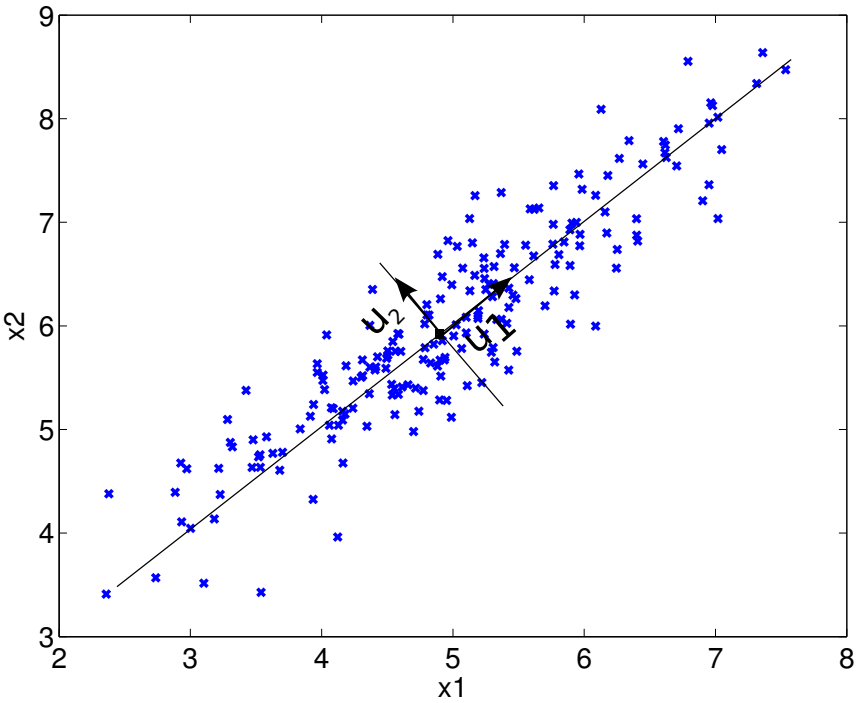
Consider the sample mean vector and the sample covariance matrix : $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$.
$\Rightarrow$ The variance of the projected data is therefore given by the scalar :

$$v(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i - \mathbf{u}^T\overline{\mathbf{x}})(\mathbf{u}^T\mathbf{x}_i - \mathbf{u}^T\overline{\mathbf{x}})^T = \mathbf{u}^T\mathbf{S}\mathbf{u} \cdot \quad (1)$$

The **principal axes** (the direction vectors) are then given by :

$$\mathbf{u} = \arg\max_{\mathbf{u}\in\mathbb{R}^d} \mathbf{u}^T\mathbf{S}\mathbf{u} \quad (2)$$

subject to $\mathbf{u}^T\mathbf{u} = 1$ and $\mathbf{u}_j^T\mathbf{u}_k = 0$ for $j \neq k$


9
8
7
6
5
4
3
2
3
4
5
6
7
8
x1
x2
u_2
u_1

Disadvantage :

The absence of a probability density model and associated likelihood measure.
Deriving PCA from the perspective of density estimation would offer a number of important advantages, including the following :

- The likelihood measure allows comparison with other density models
- We can derive EM for PCA and hence deal with possible missing values
- Possibility to perform Bayesian inference (e.g. for model selection)
- Possibility of computing the the posterior class probabilities if PCA is used to model the class-conditional densities in classification,
- The value of the probability density function would give a measure of the novelty of a new data point.
- PCA model could be extended to a mixture framework.

$\Rightarrow$ Use Probabilistic Principal Component Analysis (PPCA)

(PPCA)

université PARIS-SACLAY

Latent variable model for PPCA [Tipping and Bishop, 1997, 1999, Roweis, 1998] :

$$
\begin{aligned}
\mathbf{x}_i &= \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \text{ Observed data} = \text{linear transf. of } \mathbf{z} + \text{additive Gaussian noise} \\
\mathbf{z}_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \text{ latent variables of the principal component subspace} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ zero-mean Gaussian noise}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{x}_i|\mathbf{z}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \text{ conditional density for the observed data} \\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \text{ marginal density for the observed data}
\end{aligned}
$$

$\hookrightarrow (\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$: Parameter estimation using EM [Tipping and Bishop, 1997, 1999, Roweis, 1998]

EM for PPCA

NB : for $\boldsymbol{\mu}$, we get its closed form solution : $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$
Only $\mathbf{W}$ and σ^2 are computed in an iterative way by EM

1. **E-step** : By using the old parameters values, compute

$$\mathbb{E}[\mathbf{z}_i] = (\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^T(\mathbf{x}_i - \overline{\mathbf{x}}) \quad (3)$$

$$\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T] = \sigma^2(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1} + \mathbb{E}[\mathbf{z}_i]\mathbb{E}[\mathbf{z}_i]^T \quad (4)$$

2. **M-step**

$$\mathbf{W}_{\text{new}} = \left[\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_i]^T\right]\left[\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\right]^{-1} \quad (5)$$

$$\sigma^2_{\text{new}} = \frac{1}{nd}\sum_{i=1}^{n}\left\{\|\mathbf{x}_i - \overline{\mathbf{x}}\|^2 - 2\mathbb{E}[\mathbf{z}_i]^T\mathbf{W}_{\text{new}}^T(\mathbf{x}_i - \overline{\mathbf{x}}) + trace(\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\mathbf{W}_{\text{new}}\mathbf{W}_{\text{new}}^T\right\} \quad (6)$$

NB. Here $\mathbb{E}[.]$ is actually $\mathbb{E}[.|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}_{\text{old}}]$

Factor Analysis (FA) I

université PARIS-SACLAY

Factor Analysis (FA) [Spearman, 1904, Thurstone, 1947]
FA is closely related to PPCA
The only difference is

$$\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) \text{ conditional density for the observed data}$$

$\boldsymbol{\Psi}$ is a $d \times d$ digonal matrix ; rather than

$$\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \text{ conditional density for the observed data}$$

(isotropic covariance matrix).

Factor Analysis (FA) II

université PARIS-SACLAY

Generative model

$$
\begin{aligned}
\mathbf{x}_i &= \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \text{ Observed data} = \text{linear transf. of } \mathbf{z} + \text{additive Gaussian noise} \\
\mathbf{z}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \text{ latent variables of the principal component subspace} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ zero-mean Gaussian noise}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{x}_i|\mathbf{z}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) \text{ conditional density for the observed data} \\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}) \text{ marginal density for the observed data}
\end{aligned}
$$

EM for Facotr Analysis

université PARIS-SACLAY

1. **E-step**

$$\mathbb{E}[\mathbf{z}_i] = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{\Psi}^{-1}\mathbf{x}_i - \overline{\mathbf{x}}) \quad (7)$$

$$\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T] = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1} + \mathbb{E}[\mathbf{z}_i]\mathbb{E}[\mathbf{z}_i]^T \quad (8)$$

2. **M-step**

$$\mathbf{W}_{\text{new}} = \left[\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_i]^T\right]\left[\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\right]^{-1} \quad (9)$$

$$\mathbf{\Psi}_{\text{new}} = \text{diag}\left\{\mathbf{S} - \mathbf{W}_{\text{new}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i](\mathbf{x}_i - \overline{\mathbf{x}})^T\right\} \quad (10)$$

NB. Here $\mathbb{E}[.]$ is actually $\mathbb{E}[.|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \mathbf{\Psi}\}_{\text{old}}]$

reduction

université PARIS-SACLAY

- Introduction
- Principal Component Analysis (PCA)
- Probabilistic Principal Component Analysis (PPCA)
- Factor Analysis (FA)
- tSNE

Introduction I

université PARIS-SACLAY

Until now we have considered discrete latent data models (GMM, HMM) and continuous ones (GTM, etc) ; now we will see other continuous latent data models

The aim here is the dimensionality reduction for preprocessing, compression, feature extraction, visualization etc of high dimensional data

Principal Component Analysis (PCA) [Pearson, 1901, Hotelling, 1933],

Factor Analysis (FA)[Spearman, 1904, Thurstone, 1947],

Independent Component Analysis (ICA)[Comon, 1994, Hyvärinen, 2001], etc

are examples of well-known techniques which can be used to achieve this task.

Introduction II

PCA is a well-established technique for dimensionality reduction,

A linear projection technique that **maximizes the variance** in the projected space

Equivalently, it **minimizes the reconstruction error** (after the dimensionality reduction)

Principal Component Analysis (PCA) I

université PARIS-SACLAY

Two views of PCA :

1. First view : PCA maximizing the projected variance [Hotelling, 1933]
2. Second view : minimizing the reconstruction error (after the dimensionality reduction)

The most common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space [Hotelling, 1933].

Consider a set of observed d-dimensional data vector $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$

$\Rightarrow$ The aim is to project the data onto a space having dimensionality $M < d$ while maximizing the variance of the projected data.

Principal Component Analysis (PCA) II

université PARIS-SACLAY

Consider the sample mean vector and the sample covariance matrix :

$$\begin{aligned} \bar{\mathbf{x}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \\ \mathbf{S} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T. \end{aligned}$$

Let us first consider the projection onto a one-dimensional space ($M=1$). The direction of this space can be defined using a d-dimensional direction unit vector $\mathbf{u}_1$ (with $\mathbf{u}_1^T\mathbf{u}_1=1$)

The linear projection of a data vector $\mathbf{x}_i$ on the projected space is given by the scalar :

$$\mathbf{u}_1^T\mathbf{x}_i$$

Principal Component Analysis (PCA) III

université PARIS-SACLAY

$\Rightarrow$ The variance of the projected data is therefore given by the scalar :

$$v(\mathbf{u}_1) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}_1^T\mathbf{x}_i - \mathbf{u}_1^T\bar{\mathbf{x}})(\mathbf{u}_1^T\mathbf{x}_i - \mathbf{u}_1^T\bar{\mathbf{x}})^T = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 \cdot \quad (11)$$

The **principal axe** (the direction vector) is then given by :

$$\mathbf{u}_1 = \arg\max_{\mathbf{u}_1 \in \mathbb{R}^d} \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 \text{ subject to } \mathbf{u}_1^T\mathbf{u}_1 = 1 \cdot \quad (12)$$

The normalisation condition is namely to prevent $\|\mathbf{u}_1\| \to \infty$

This is a constrained maximization problem $\Rightarrow$ Use a Lagrange multiplier to solve it

Principal Component Analysis (PCA) IV

université PARIS-SACLAY

The unconstrained maximization is therefore given by

$$\mathbf{u} = \arg\max_{\mathbf{u}_1 \in \mathbb{R}^d} \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1) \cdot \tag{13}$$

$$\frac{\partial \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda (1 - \mathbf{u}_1^T \mathbf{u}_1)}{\partial \mathbf{u}_1} = 0 \quad \overset{\frac{\partial \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1}{\partial \mathbf{u}_1} = (\mathbf{S} + \mathbf{S}^T)\mathbf{u}}{\Leftrightarrow} \quad 2\mathbf{S}\mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

$$\Leftrightarrow \quad \mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \tag{14}$$

$\Rightarrow$ $\mathbf{u}_1$ must be an eigenvector of the data covariance matrix S, with eigenvalue λ_1

The variance (11) in the projected space is then given by

$$\begin{aligned} \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 &= \mathbf{u}^T \lambda_1 \mathbf{u}_1 \\ &= \lambda_1 \quad (\text{ since } \mathbf{u}_1^T \mathbf{u}_1 = 1) \end{aligned} \tag{15}$$

Principal Component Analysis (PCA) V

université PARIS-SACLAY

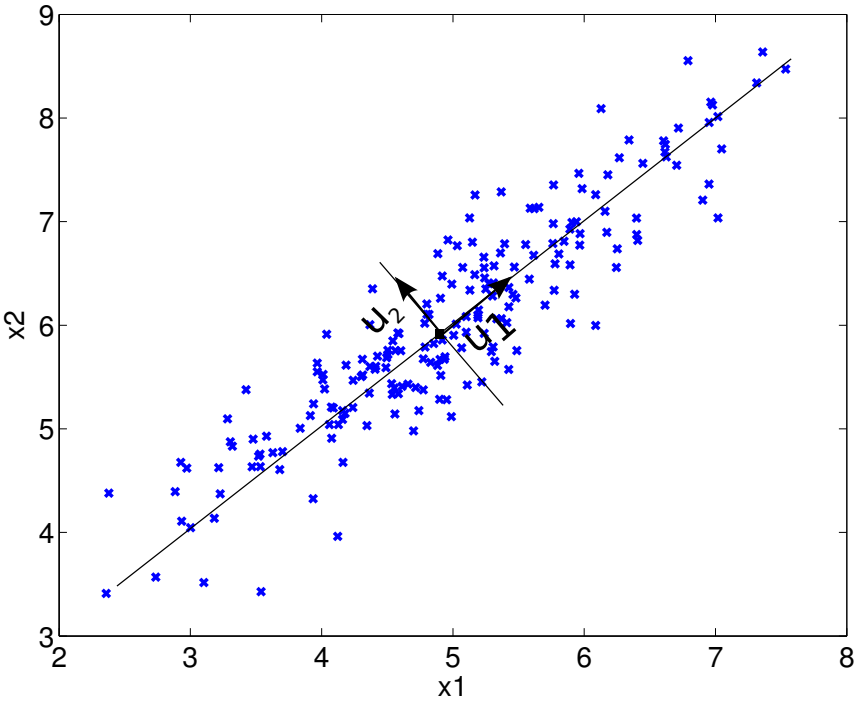
$\Rightarrow$ The variance will be maximum if we set $\mathbf{u}_1$ equal to the eigenvector having the **largest** eigenvalue λ_1.

$\Rightarrow$ The eigenvector $\mathbf{u}_1$ is known as the first **principal component** or the first principal axe.

We can define additional principal components in an incremental way :

The new direction to be chosen is that which maximizes the projected variance amongst all possible directions orthogonal to those already considered.

Principal Component Analysis (PCA) VI



Principal Component Analysis (PCA) VII

université PARIS-SACLAY

General case :

Consider the general case of a projection space of dimension M (M principal components).

The optimal linear projection for which the projected variance is maximized is defined by M eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_M$ of the data covariance matrix $\mathbf{S}$ corresponding to the largest eigenvalues $\lambda_1, \ldots, \lambda_M$

The result was shown for one principal components ($M = 1$).

Now suppose that the result holds for M principal components and we aim to show that it holds for $M + 1$ (by induction)

For $\mathbf{u}_{M+1}$, the projected variance in the direction $\mathbf{u}_{M+1}$ is given by

$$\mathbf{u}_{M+1}^T \mathbf{S} \mathbf{u}_{M+1}$$

Principal Component Analysis (PCA) VIII

université PARIS-SACLAY

We therefore maximize the variance $\mathbf{u}_{M+1}^T \mathbf{S} \mathbf{u}_{M+1}$ by taking into account the normalization constraint and the orthogonality constraints :

- normalization constraint : $\mathbf{u}_{M+1}^T$ is normalized to unit length, that is : $\mathbf{u}_{M+1}^T \mathbf{u}_{M+1} = 1$
- orthogonal constraint : $\mathbf{u}_{M+1}$ orthogonal to the existing vectors $\mathbf{u}_1, \ldots, \mathbf{u}_M$, that is : $\mathbf{u}_{M+1}^T \mathbf{u}_k = 0$ for $k \neq M+1$

$\Rightarrow$ Use a Lagrange multiplier λ_{M+1} and Lagrange multipliers $\eta_k, k = 1, \ldots, M$ to enforce these constraints.

Principal Component Analysis (PCA) IX

université PARIS-SACLAY

$\Rightarrow$ Thus we solve the following unconstrained maximization problem

$$\begin{aligned}\mathbf{u}_{M+1} &= \arg\max_{\mathbf{u}_{M+1}} \mathbf{u}_{M+1}^T \mathbf{S}\mathbf{u}_{M+1} + \lambda_{M+1}(1 - \mathbf{u}_{M+1}^T\mathbf{u}_{M+1}) + \sum_{k=1}^{m} \eta_k \mathbf{u}_{M+1}^T \mathbf{u}_k \\ &= \arg\max_{\mathbf{u}_{M+1}} v\left(\{\mathbf{u}_k, \eta_k\}_{k=1}^m, \lambda_{M+1}, \mathbf{u}_{M+1}\right) \qquad (16)\end{aligned}$$

$$\frac{\partial v\left(\{\mathbf{u}_k, \eta_k\}_{k=1}^m, \lambda_{M+1}, \mathbf{u}_{M+1}\right)}{\partial \mathbf{u}_{M+1}} = 0 \Leftrightarrow 2\mathbf{S}\mathbf{u}_{M+1} - 2\lambda_{M+1}\mathbf{u}_{M+1} + \sum_{k=1}^{m} \eta_k \mathbf{u}_k = 0$$

$$\Leftrightarrow \underbrace{\mathbf{u}_j^T \mathbf{S}\mathbf{u}_{M+1}}_{0} - \lambda_{M+1} \underbrace{\mathbf{u}_j^T \mathbf{u}_{M+1}}_{0} + \underbrace{\sum_{k=1}^{m} \eta_k \mathbf{u}_j^T \mathbf{u}_k}_{\eta_j \mathbf{u}_j^T \mathbf{u}_j = \eta_j} = 0$$

$$\Leftrightarrow \eta_j = 0 \quad \text{for} \quad j = 1, \ldots, m \qquad (17)$$

We therefore obtain : $\mathbf{S}\mathbf{u}_{M+1} = \lambda_{M+1}\mathbf{u}_{M+1}$

Principal Component Analysis (PCA) X

université PARIS-SACLAY

$\Rightarrow \mathbf{u}_{M+1}$ must be an eigenvector of $\mathbf{S}$ with eigenvalue λ_{M+1} .

The projected variance in direction $\mathbf{u}_{M+1}$ is therefore given by

$$\begin{aligned} \mathbf{u}_{M+1}^T \mathbf{S} \mathbf{u}_{M+1} &= \mathbf{u}_{M+1}^T \lambda_{M+1} \mathbf{u}_{M+1} \\ &= \lambda_{M+1} \end{aligned} \tag{18}$$

Principal Component Analysis (PCA) XI

université PARIS-SACLAY

$\Rightarrow$ The variance will be maximum if we set $\mathbf{u}_{M+1}$ equal to the eigenvector having the **largest** eigenvalue λ_{M+1} amongst those not previously selected.

Thus the result holds also for projection spaces of dimensionality $M+1$, which completes the inductive step.

Since we have already shown this result explicitly for $M=1$, if follows that the result must hold for any $M \leq d$.

Principal Component Analysis (PCA) XII

université PARIS-SACLAY

Second view : minimization of the reconstruction error PCA minimizes the reconstruction error, that is the squared error between a data point $\mathbf{x}_i$ and its approximation $\tilde{\mathbf{x}}_i$, averaged over all the data points : $J = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2$

Consider one direction $\mathbf{u}$ in the projection space

Here we will show minimizing this error w.r.t $\mathbf{u}$ is equivalent to maximizing the projected variance (11) on the direction $\mathbf{u}$

Let us assume that all the original vectors $\mathbf{x}_i$ have been centered :

$$\mathbf{x}_i^c = \mathbf{x}_i - \bar{\mathbf{x}}_i$$

By using the fact that the projection of a data vector $\mathbf{x}$ onto the direction $\mathbf{u}$ is given by the scalar $\mathbf{u}^T\mathbf{x}$; $\mathbf{x}$ is then represented by $(\mathbf{u}^T\mathbf{x})\mathbf{u}$ in the projected space,

Principal Component Analysis (PCA) XIII

université PARIS-SACLAY

The reconstruction error for the centred data is then given by :

$$
\begin{aligned}
J(\mathbf{u}) &= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i^c - \tilde{\mathbf{x}}_i^c\|^2 \frac{1}{n}\sum_{i=1}^{n}\|(\mathbf{x}_i - \bar{\mathbf{x}}_i) - [\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)\mathbf{u}]\|^2 \\
&= \frac{1}{n}\Big(\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|^2 - 2\sum_{i=1}^{n}[\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)][\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)] + \sum_{i=1}^{n}[\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)]^2\|\mathbf{u}\|^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|^2 - \frac{2}{n}\sum_{i=1}^{n}[\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)]^2 + \frac{1}{n}\sum_{i=1}^{n}[\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}[\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_i - \bar{\mathbf{x}}_i)\mathbf{u} \\
&= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|^2 - \mathbf{u}^T\mathbf{S}\mathbf{u}
\end{aligned}
\tag{19}
$$

The first term does not depend on $\mathbf{u}$. Thus the vector $\mathbf{u}$ that minimizes $J(\mathbf{u})$ is the same one that maximizes the projected variance $\mathbf{u}^T\mathbf{S}\mathbf{u}$ (11).

Principal Component Analysis (PCA) XIV

université PARIS-SACLAY

Summary :

PCA reduces the dimensionality of the data while retaining as much as possible of the variation present in the original dataset

$$\mathbf{X} : [n \times d\ ] \xrightarrow{\text{Linear Proj. onto } \mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_M]} \tilde{\mathbf{X}} = \mathbf{X}_c\mathbf{U} : [n \times M]; M \leq d$$

To perform PCA on a data set $\mathbf{X}$

1. calculate the mean data vector $\bar{\mathbf{x}}$
2. calculate the data covariance matrix $\mathbf{S}$
3. calculate the eigenvectors and the corresponding eigenvalues of $\mathbf{S}$ (e.g., by using the `eig` function in Matlab)

Principal Component Analysis (PCA) XV

université PARIS-SACLAY

4. the eigenvalues $\lambda_1, \ldots, \lambda_d$ are sorted in decreasing order ; the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are placed according to the resulting order

5. the projection space (the space of principal axes) is then obtained by taking the M first eigenvectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]; M \leq d$
6. the projected data are given by $\tilde{\mathbf{X}} = \mathbf{X}_c \mathbf{U}$ where $\mathbf{X}_c$ is the centered data matrix

How to choose M ? for example one way is to choose the first M components that capture a specified percentage e.g., 90%, 95%, or 99%, of the cumulative percentage of variance. $cpv(M) = 100 \left(\frac{\sum_{m=1}^{M} \lambda_m}{\sum_{m=1}^{d} \lambda_m} \right) \%$

Disadvantage : One disadvantage of both these definitions of PCA is the absence of a probability density model and associated likelihood measure.

Principal Component Analysis (PCA) XVI

université PARIS-SACLAY

Deriving PCA from the perspective of density estimation would offer a number of important advantages, including the following :

- The likelihood measure would permit comparison with other density models
- We can derive EM for PCA and hence deal with missing values in the data set
- Possibility to perform Bayesian inference (e.g. for model selection)
- Possibility of computing the the posterior class probabilities if PCA is used to model the class-conditional densities in a classification problem,
- The value of the probability density function would give a measure of the novelty of a new data point.
- PCA model could be extended to a mixture framework.

$\Rightarrow$ Use Probabilistic Principal Component Analysis (PPCA)

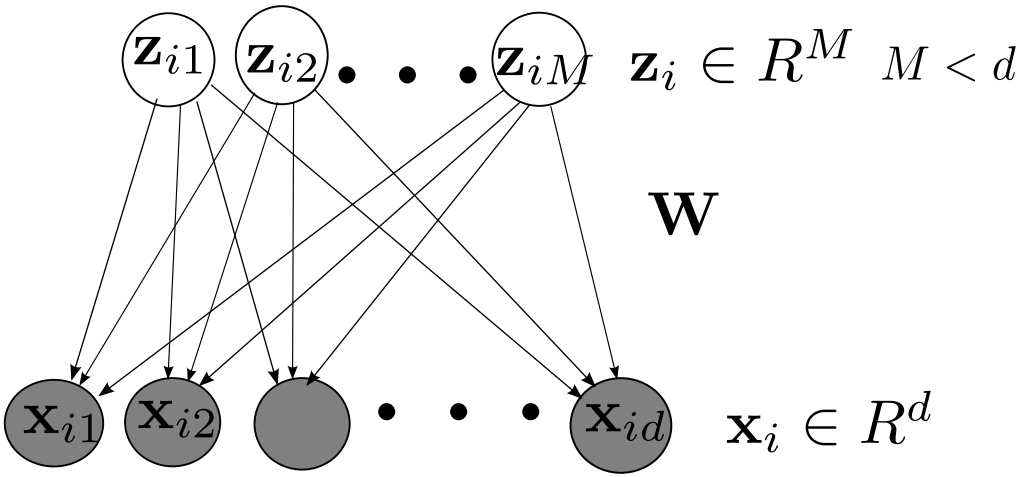
PCA can be formulated into a probabilistic framework : the Probabilistic Principal Component Analysis (PPCA) [Tipping and Bishop, 1997, 1999, Roweis, 1998]

The PC can be expressed as the maximum likelihood solution of a latent **continuous** variable model and the model parameter are optimized using EM [Tipping and Bishop, 1997, 1999, Roweis, 1998]

$\Rightarrow$ Generative formulation : **the latent variable model** for PPCA :

$$
\begin{aligned}
\mathbf{x}_i &= \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \text{ Observed data = linear transf. of } \mathbf{z} + \text{additive Gaussian noise} \\
\mathbf{z}_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \text{ latent variables of the principal component subspace} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ zero-mean Gaussian noise} \\
\mathbf{x}_i|\mathbf{z}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \text{ conditional density for the observed data} \\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \text{ marginal density for the observed data}
\end{aligned}
\tag{20}
$$

Probabilistic Principal Component Analysis (PPCA) II



Maximum Likelihood for PPCA I

Model parameters : $(\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$ where $\mathbf{W}$ a $[d \times M]$ matrix whose columns represent the principal subspace, $\boldsymbol{\mu}$ a d-dimensional vector

- Assume we have an i.i.d sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
- The observed-data log-likelihood is given by :

$$\begin{aligned}
\mathcal{L}(\mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= \log \prod_{i=1}^{n} p(\mathbf{x}_i; \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \log \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{\mu}, \underbrace{\mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}}_{\mathbf{C}}) \\
&= -\frac{nd}{2} \log 2\pi - \frac{n}{2} \log |\mathbf{C}| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\
&= -\frac{n}{2} \left(d \log 2\pi + \frac{1}{2} \log |\mathbf{C}| + \text{trace} \left\{ \mathbf{C}^{-1} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \right) \\
&= -\frac{n}{2} \left(d \log 2\pi + \frac{1}{2} \log |\mathbf{C}| + \text{trace} \left\{ \mathbf{C}^{-1} \mathbf{S} \right\} \right)
\end{aligned} \tag{21}$$

Maximum Likelihood for PPCA II

université PARIS-SACLAY

$\Rightarrow$ Analytical solutions

Maximum Likelihood for PPCA III

ML estimates [Tipping and Bishop, 1999] :

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \bar{\mathbf{x}} \tag{22}$$

$$\hat{\sigma}^2 = \frac{1}{d-M}\sum_{m=M+1}^{d}\lambda_m \tag{23}$$

$$\hat{\mathbf{W}} = \mathbf{U}_M(\mathbf{L}_M - \hat{\sigma^2}\mathbf{I})^{1/2}\mathbf{R}, \tag{24}$$

where $\mathbf{U}_M$ is a $[d \times M]$ matrix whose columns are the first M eigenvectors $[\mathbf{u}_1, \ldots, \mathbf{u}_M]$ of the data covariance-matrix $\mathbf{S}$ corresponding to the the first M largest eigenvalues $[\lambda_1, \ldots, \lambda_M]$

$\mathbf{L}_M$ is an $[M \times M]$ matrix whose diagonal elements are the corresponding eigenvalues $[\lambda_1, \ldots, \lambda_M]$

$\mathbf{R}$ is an $[M \times M]$ arbitrary orthogonal matrix $(\mathbf{R}\mathbf{R}^T = \mathbf{I})$

EM for PPCA I

université PARIS-SACLAY

The PPCA is expressed as a latent data model : so we can use EM to find the ML estimates for PPCA

While we have exact solutions ; using EM, as it is iterative, may have an advantage in spaces of high dimensionality compared to when working with the sample data covariance matrix $\mathbf{S}$ (for the eignvalues and eigenvalues)

The EM procedure can also be extended to Factor Analysis for which there is no analytical solutions

The log-likelihood of the PPCA model parameters $(\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$ for the complete-data $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}_1, \mathbf{z}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$:

$$\mathcal{L}_c(\mathbf{W}, \boldsymbol{\mu}, \sigma^2; \mathbf{X}, \mathbf{Z}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i; \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \sum_{i=1}^{n} [\log p(\mathbf{x}_i|\mathbf{z}_i) + \log p(\mathbf{z}_i)]$$

EM for PPCA II

université PARIS-SACLAY

Complete-data log-likelihood :

$$\begin{aligned}
\mathcal{L}_c(\mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= \sum_{i=1}^{n} [\log p(\mathbf{x}_i|\mathbf{z}_i) + \log p(\mathbf{z}_i)] \\
&= \sum_{i=1}^{n} [\log \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}) + \log \mathcal{N}(\mathbf{z}_i; \mathbf{0}, \sigma^2\mathbf{I})] \\
&= -\sum_{i=1}^{n} \Big\{ \frac{d}{2}\log(2\pi\sigma^2) + \frac{1}{2} trace(\mathbf{z}_i\mathbf{z}_i^T) + \frac{1}{2\sigma^2}\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \\
&\quad - \frac{1}{\sigma^2}\mathbf{z}_i^T\mathbf{W}^T(\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2\sigma^2} trace(\mathbf{z}_i\mathbf{z}_i^T\mathbf{W}\mathbf{W}^T \Big\} \qquad (25)
\end{aligned}$$

Expected complete-data log-likelihood (the Q-function) :

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_c(\mathbf{W}, \boldsymbol{\mu}, \sigma^2)|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}_{\text{old}}] = -\sum_{i=1}^{n} \Big\{ \frac{d}{2}\log(2\pi\sigma^2) + \frac{1}{2} trace(\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]) + \frac{1}{2\sigma^2}\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \\
- \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_i]^T\mathbf{W}^T(\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2\sigma^2} trace(\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\mathbf{W}\mathbf{W}^T \Big\} \qquad (26)
\end{aligned}$$

EM for PPCA III

NB : for $\boldsymbol{\mu}$, we get its closed form solution : $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$

Only $\mathbf{W}$ and σ^2 are computed in an iterative way by EM

1. **E-step** : By using the old parameters values, compute

$$\mathbb{E}[\mathbf{z}_i] = (\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^T(\mathbf{x}_i - \bar{\mathbf{x}}) \quad (27)$$

$$\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T] = \sigma^2(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1} + \mathbb{E}[\mathbf{z}_i]\mathbb{E}[\mathbf{z}_i]^T \quad (28)$$

2. **M-step**

$$\mathbf{W}_{\text{new}} = \left[\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\mathbb{E}[\mathbf{z}_i]^T\right]\left[\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\right]^{-1} \quad (29)$$

$$\sigma^2_{\text{new}} = \frac{1}{nd}\sum_{i=1}^{n}\left\{\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 - 2\mathbb{E}[\mathbf{z}_i]^T\mathbf{W}_{\text{new}}^T(\mathbf{x}_i - \bar{\mathbf{x}}) + trace(\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\mathbf{W}_{\text{new}}\mathbf{W}_{\text{new}}^T\right\} \quad (30)$$

NB. Here $\mathbb{E}[.]$ is actually $\mathbb{E}[.|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}_{\text{old}}]$

Factor Analysis (FA) I

université PARIS-SACLAY

Factor Analysis (FA) [Spearman, 1904, Thurstone, 1947]

FA is closely related to PPCA

The only difference is

$\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})$ conditional density for the observed data

$\boldsymbol{\Psi}$ is a $d \times d$ digonal matrix ; rather than

$\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I})$ conditional density for the observed data

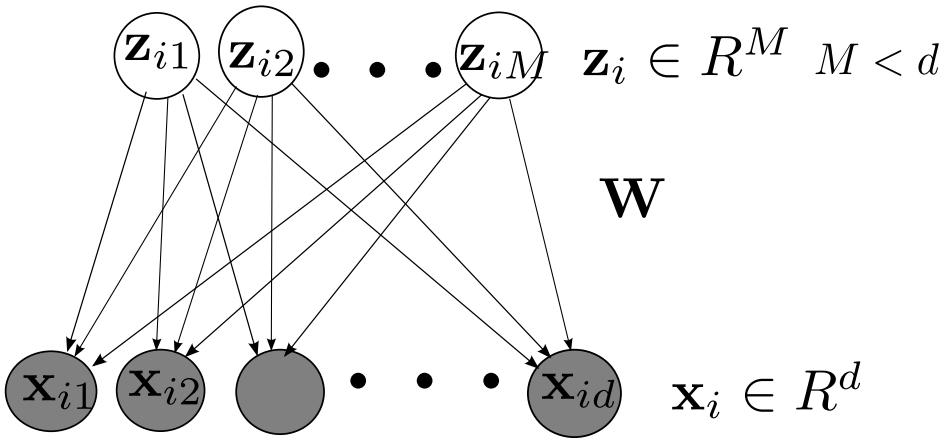
(isotropic covariance matrix).

Factor Analysis (FA) II

université PARIS-SACLAY

Generative model

$$
\begin{aligned}
\mathbf{x}_i &= \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \text{ Observed data} = \text{linear transf. of } \mathbf{z} + \text{additive Gaussian noise} \\
\mathbf{z}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \text{ latent variables of the principal component subspace} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ zero-mean Gaussian noise} \\
\mathbf{x}_i|\mathbf{z}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) \text{ conditional density for the observed data} \\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}) \text{ marginal density for the observed data}
\end{aligned}
$$



Factor Analysis (FA) III

université PARIS-SACLAY

EM

1 E-step

$$\mathbb{E}[\mathbf{z}_i] = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{\Psi}^{-1}\mathbf{x}_i - \bar{\mathbf{x}}) \quad (31)$$

$$\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T] = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1} + \mathbb{E}[\mathbf{z}_i]\mathbb{E}[\mathbf{z}_i]^T \quad (32)$$

2 M-step

$$\mathbf{W}_{\text{new}} = \left[\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\mathbb{E}[\mathbf{z}_i]^T\right]\left[\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\right]^{-1} \quad (33)$$

$$\mathbf{\Psi}_{\text{new}} = \text{diag}\left\{\mathbf{S} - \mathbf{W}_{\text{new}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i](\mathbf{x}_i - \bar{\mathbf{x}})^T\right\} \quad (34)$$

NB. Here $\mathbb{E}[.]$ is actually $\mathbb{E}[.|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \mathbf{\Psi}\}_{\text{old}}]$

Illustration on PCA (Face Recognition) seen in classroom

tSNE : course materials and pdf available on

https://chamroukhi.com/Teaching/ML-MscAI-DS/tSNE-en.pdf
https://chamroukhi.com/Teaching/ML-MscAI-DS/tSNE-fr.pdf

Some References I

P. Comon. Independent Component Analysis, a new concept ? *Signal Processing*, 36(3) :287–314, 1994. Special issue on Higher-Order Statistics.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 : 417–441, 1933.

A. Hyvärinen. *Independant Component Analysis*. Wiley, 2001.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901.

S. Roweis. EM algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Proceedings of the 11th Conference on Advances in Neural Information Processing Systems (NIPS)*, volume 10. MIT Press, 1998.

C. Spearman. General intelligence, objectively determined and measured. *American Journal of psychology*, 15 :201–293, 1904.

L. L. Thurstone. *Multiple Factor Analysis*. University of Chicago Press, 1947.

M. E. Tipping and C. Bishop. Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Comuting Research Group, Aston University, 1997.

M. E. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61 : 611–622, 1999.