Probabilistic dimensionality reduction

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"
a.a. 2024-2025

Giorgio Gambosi

1 Factor Analysis

Factor analysis is one of the simplest and most fundamental generative latent models, the first one we consider here where both the observed variable $\mathbf{x}$ and the latent variable $\mathbf{z}$ are real. At the same time, the model is also simple enough to make it possible to make it feasible to compute the conditional probability $p(\mathbf{z}|\mathbf{x})$.

In particular, the model assume that each element $\bar{\mathbf{x}}_i \in \mathbb{R}^D$ in the observable dataset is related to the value of a latent variable (also called a factor here) $\bar{\mathbf{z}}_i \in \mathbb{R}^d$ through:

- a linear projection from the d-dimensional space $\mathbb{R}^d$ of $\mathbf{z}$ to the D-dimensional space $\mathbb{R}^D$ of $\mathbf{x}$
- a translation of the result within $\mathbb{R}^D$
- an additional (smaller) random translation within $\mathbb{R}^D$

This is specified by the equation

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where (see Figure 1)

- $\mathbf{z} \in \mathbb{R}^d$ is a latent variable whose distribution is assumed gaussian with 0 mean and unitary covariance matrix: hence $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$
- $\mathbf{W} \in \mathbb{R}^{D \times d}$ is a linear projection of any point in $\mathbb{R}^d$ to a point in $\mathbb{R}^D$
- $\boldsymbol{\mu} \in \mathbb{R}^D$ is a translation of points in $\mathbb{R}^D$
- $\boldsymbol{\epsilon} \in \mathbb{R}^D$ is a gaussian noise for the final random translation: noise covariance on different dimensions is assumed to be 0. That is, its distribution is $\mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} \in \mathbb{R}^{D \times D}$ is a diagonal matrix, with $\boldsymbol{\Psi}_{ii}$ the noise variance along the i-th dimension.

Background on Multivariate Gaussian Distribution

Let us consider the following situation, where $\mathbf{x}$ and $\mathbf{z}$ are two random variables:

1. $\mathbf{z}$ is normally distributed $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{z}, \Sigma_\mathbf{z})$
2. there exist $\mathbf{A} \in \mathbb{R}^{D \times d}$, $\mathbf{b} \in \mathbb{R}^D$ such that the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ is a gaussian $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{A}\mathbf{z} + \mathbf{b}, \Sigma_{\mathbf{x}\mathbf{z}})$

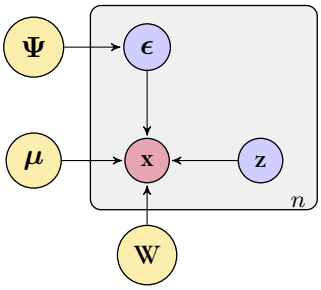

Figure 1: The latent variables $\boldsymbol{\epsilon}$ and $\mathbf{z}$ are normally distributed on the observed and the latent space, respectively: they can be both seen as random noise $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi})$ and $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. The observed variable $\mathbf{x}$ is deterministically dependent from them as $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$. However, a probabilistic dependence from $\mathbf{z}$ alone can be expressed through the conditional distribution $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{z}; \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$.

this is denoted as **linear gaussian model** and, in this framework, both the marginal distribution of $\mathbf{x}$ and the inverse conditional distribution of $\mathbf{z}|\mathbf{x}$ are also Gaussian. In particular

- For the marginal distribution, $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \Sigma_\mathbf{x})$, with

$$\begin{aligned}\boldsymbol{\mu}_\mathbf{x} &= \mathbf{A}\boldsymbol{\mu}_\mathbf{z} + \mathbf{b} \\ \Sigma_\mathbf{x} &= \Sigma_{\mathbf{xz}} + \mathbf{A}\Sigma_\mathbf{z}\mathbf{A}^T\end{aligned}$$

- For the conditional distribution, $\mathbf{z}|\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$, with

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} &= (\Sigma_\mathbf{z}{}^{-1} + \mathbf{A}^T\Sigma_{\mathbf{xz}}{}^{-1}\mathbf{A})^{-1}(\mathbf{A}^T\Sigma_{\mathbf{xz}}^{-1}(\mathbf{x} - \mathbf{b}) + \Sigma_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{x}) \\ \Sigma_{\mathbf{z}|\mathbf{x}} &= (\Sigma_\mathbf{z}{}^{-1} + \mathbf{A}^T\Sigma_{\mathbf{xz}}^{-1}\mathbf{A})^{-1}\end{aligned}$$

The Factor Analysis Model

As already stated, the prior distribution of the latent variable is assumed to be a multivariate Gaussian distribution.

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$

and the observed value $\mathbf{x}$ is obtained from $\mathbf{z}$ through

1. the linear projection of $\mathbf{z}$ by $\mathbf{W} \in \mathbb{R}^{D\times d}$,
2. applying some linear translation $\boldsymbol{\mu} \in \mathbb{R}^D$, and
3. adding a Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^D$ with mean $\mathbf{0}$ and covariance $\boldsymbol{\Psi} \in \mathbb{R}^{D\times D}$.

As a consequence, the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ is

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

Factor Analysis is then a linear Gaussian model with $\boldsymbol{\mu}_z = \mathbf{0}$, $\Sigma_z = \mathbf{I}$, $\mathbf{A} = \mathbf{W}$, $\mathbf{b} = \boldsymbol{\mu}$, $\Sigma_{\mathbf{x}|\mathbf{z}} = \boldsymbol{\Psi}$. By applying its properties, we get:

- the marginal distribution, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi})$

- the inverse conditional distribution, $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{x}; \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$, with

$$\Sigma_{\mathbf{z}|\mathbf{x}} = (\mathbf{I} + \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W})^{-1} \triangleq \mathbf{G} \in \mathbb{R}^{d \times d}$$
$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = \mathbf{G}\mathbf{W}^T \boldsymbol{\Psi}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \in \mathbb{R}^d$$

This distribution can be exploited to map points onto the latent space. In particular, the conditional expectation

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = \mathbf{G}\mathbf{W}^T \boldsymbol{\Psi}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \in \mathbb{R}^d$$

can be assumed as the point in latent space corresponding to $\mathbf{x} \in \mathbb{R}^D$.

Maximization of likelihood in FA

The log-likelihood of the observed dataset in the model is

$$\begin{aligned}\log p(\mathrm{X}|\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}) &= \sum_{i=1}^{n} \log p(\bar{\mathrm{x}}_i|\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \mathcal{N}(\bar{\mathrm{x}}_i; \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^T) \\ &= -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log|\boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^T| - \frac{1}{2} \sum_{i=1}^{n} (\bar{\mathrm{x}}_i - \boldsymbol{\mu})(\boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^T)^{-1} (\bar{\mathrm{x}}_i - \boldsymbol{\mu})^T\end{aligned}$$

Setting the gradient wrt $\boldsymbol{\mu}$ to 0 results into

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathrm{x}}_i \triangleq \bar{\mathbf{x}} \in \mathbb{R}^D$$

However, no closed form solution for $\mathbf{W}$ and $\boldsymbol{\Psi}$ can be obtained by setting the corresponding gradients to $\mathbf{0}$. Iterative techniques such as EM can then be applied to maximize the log-likelihood with respect to these parameters.

Expectation-Maximization for FA

By definition, the algorithm operates by alternatively computing (in the E-step)

$$p(\mathrm{Z}|\mathrm{X}; \boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} p(\mathbf{z}|\bar{\mathrm{x}}_i; \boldsymbol{\theta}^{(k)})$$

given the parameter value $\boldsymbol{\theta}^{(k)}$ and then (in the M-step) maximizing

$$\mathop{\mathbb{E}}_{p(\mathrm{Z}|\mathrm{X}; \boldsymbol{\theta}^{(k)})} [\log p(\mathrm{X}, \mathrm{Z}; \boldsymbol{\theta})] = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{p(\mathbf{z}|\bar{\mathrm{x}}_i; \boldsymbol{\theta}^{(k)})} [\log p(\bar{\mathrm{x}}_i, \mathrm{z}; \boldsymbol{\theta})]$$

with respect to the parameter $\boldsymbol{\theta}$, obtaining the new value $\boldsymbol{\theta}^{(k+1)}$.

M-step Let us first observe that in the case of FA, we have $\boldsymbol{\theta} = (\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi})$.

For what regards maximization wrt $\boldsymbol{\mu}$, we already observed that the optimum value for such parameter is

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathrm{x}}_i \in \mathbb{R}^D$$

regarding maximization wrt $\mathbf{W}$ and $\mathbf{\Psi}$, we skip some technical details, stating, without proof, that

$$\mathbf{W} = \left(\sum_{i=1}^{n}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\hat{\boldsymbol{\mu}}_i^T\right)\left(\sum_{i=1}^{n}\tilde{\boldsymbol{\mu}}_i\right)^{-1} \in \mathbb{R}^{D\times d}$$
$$\mathbf{\Psi} = \mathbf{diag}\left(\mathbf{S} - \frac{1}{n}\mathbf{W}\sum_{i=1}^{n}\tilde{\boldsymbol{\mu}}_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T\right) \in \mathbb{R}^{D\times D}$$

where

1. $\hat{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{\mu}}_i$ are the expectations wrt distribution $p(\mathbf{z}|\bar{\mathbf{x}}_i; \mathbf{W}, \boldsymbol{\mu}, \mathbf{\Psi})$ of $\mathbf{z}$ and $\mathbf{z}\mathbf{z}^T$, respectively

$$\hat{\boldsymbol{\mu}}_i \triangleq \mathop{\mathbb{E}}_{p(\mathbf{z}|\bar{\mathbf{x}}_i;\mathbf{W},\boldsymbol{\mu},\mathbf{\Psi})}[\,\mathbf{z}\,] = \int_{\mathcal{Z}} \mathbf{z}p(\mathbf{z}|\bar{\mathbf{x}}_i; \mathbf{W}, \boldsymbol{\mu}, \mathbf{\Psi})d\mathbf{z} \in \mathbb{R}^d$$
$$\tilde{\boldsymbol{\mu}}_i \triangleq \mathop{\mathbb{E}}_{p(\mathbf{z}|\bar{\mathbf{x}}_i;\mathbf{W},\boldsymbol{\mu},\mathbf{\Psi})}[\,\mathbf{z}\mathbf{z}^T\,] = \int_{\mathcal{Z}} \mathbf{z}\mathbf{z}^T p(\mathbf{z}|\bar{\mathbf{x}}_i; \mathbf{W}, \boldsymbol{\mu}, \mathbf{\Psi})d\mathbf{z} \in \mathbb{R}^{d\times d}$$

2. the **diag** operator sets to 0 all non diagonal elements
3. S is the scatter matrix of $\mathbf{X}$

$$\mathbf{S} \triangleq \frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \in \mathbb{R}^{D\times D}$$

E-step The conditional expectations $\hat{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{\mu}}_i$ are computed here. They can be shown to be

$$\hat{\boldsymbol{\mu}}_i = \mathbf{G}\mathbf{W}^T\mathbf{\Psi}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$$
$$\tilde{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\mu}}_i\hat{\boldsymbol{\mu}}_i^T + \mathbf{G}$$

where, as shown above,

$$\mathbf{G} = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}$$

The EM algorithm in factor analysis is then summarized as follows. The centroid of data, $\bar{\mathbf{x}}$, is computed and, from it, all $\bar{\mathbf{x}}_i$. Then, at every step k, we iteratively solve as:

$$\begin{aligned}
&\text{for } i = 1, \ldots, n: \\
&\qquad \hat{\boldsymbol{\mu}}_i^{(k)} \leftarrow \mathbf{G}^{(k)}(\mathbf{W}^{(k)})^T(\mathbf{\Psi}^{(k)})^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}) \\
&\qquad \tilde{\boldsymbol{\mu}}_i^{(k)} \leftarrow \hat{\boldsymbol{\mu}}_i^{(k)}(\hat{\boldsymbol{\mu}}_i^{(k)})^T + \mathbf{G}^{(k)} \\
&\mathbf{W}^{(k+1)} \leftarrow \left(\sum_{i=1}^{n}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\hat{\boldsymbol{\mu}}_i^{(k)})^T\right)\left(\sum_{i=1}^{n}\tilde{\boldsymbol{\mu}}_i^{(k)}\right)^{-1} \\
&\mathbf{\Psi}^{(k+1)} \leftarrow \frac{1}{n}\mathrm{diag}\left(\mathbf{S} - \mathbf{W}^{(k+1)}\sum_{i=1}^{n}\hat{\boldsymbol{\mu}}_i^{(k)}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T\right) \\
&\mathbf{G}^{(k+1)} \leftarrow \left(\mathbf{I} + (\mathbf{W}^{(k+1)})^T(\mathbf{\Psi}^{(k+1)})^{-1}\mathbf{W}^{(k+1)}\right)^{-1}
\end{aligned}$$

until convergence.

2 Probabilistic PCA

Probabilistic PCA is defined through a simplification of the factor analysis model. In particular, all the rest being equal, the noise covariance matrix is assumed to have equal variance for all dimensions. That is,

$$\mathbf{\Psi} = \sigma^2 \mathbf{I} \in \mathbb{R}^{D \times D}$$

The resulting model is described graphically in Figure 2.

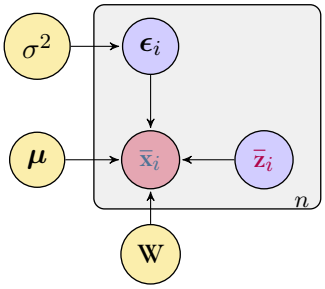

Figure 2: The latent variables $\boldsymbol{\epsilon}$ and $\mathbf{z}$ are normally distributed on the observed and the latent space, respectively: they can be both seen as random noise $p(\boldsymbol{\epsilon}; \sigma^2) = \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \sigma^2\mathbf{I})$ and $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. The observed variable $\mathbf{x}$ is deterministically dependent from them as $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$. However, a probabilistic dependence from $\mathbf{z}$ alone can be expressed through the conditional distribution $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \mathbf{I}\sigma^2)$.

Expectation-Maximization for Probabilistic PCA

Expectation maximization can be applied to maximize the log-likelihood of the observed dataset $\mathbf{X}$ wrt the parameters $\mathbf{W}$, $\boldsymbol{\mu}$, σ^2.

Being PPCA a particular case of factor analysis, the E and M steps can be derived from the ones defined for FA, substituting the new noise covariance matrix $\sigma^2\mathbf{I}$ to the more general $\mathbf{\Psi}$.

This results in the following:

$$\hat{\boldsymbol{\mu}}_i = \beta \mathbf{G} \mathbf{W}^T (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$$
$$\tilde{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^T + \mathbf{G}$$

where $\beta = \frac{1}{\sigma^2}$ is the **precision**.

It can be proved that the algorithm behaves, at each step, as follows

$$\text{for } i = 1, \dots, n:$$
$$\hat{\boldsymbol{\mu}}_i^{(k)} \leftarrow \beta^{(k)} \mathbf{G}^{(k)} \left(\mathbf{W}^{(k)}\right)^T (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$$
$$\tilde{\boldsymbol{\mu}}_i^{(k)} \leftarrow \hat{\boldsymbol{\mu}}_i^{(k)} \hat{\boldsymbol{\mu}}_i^{(k-1)T} + \mathbf{G}^{(k)}$$
$$\mathbf{W}^{(k+1)} \leftarrow \left(\sum_{i=1}^{n} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}) (\hat{\boldsymbol{\mu}}_i^{(k)})^T \right) \left(\sum_{i=1}^{n} \tilde{\boldsymbol{\mu}}_i^{(k)} \right)^{-1}$$
$$\beta^{(k+1)} \leftarrow nD \left(\sum_{i=1}^{n} ||\bar{\mathbf{x}}_i - \bar{\mathbf{x}}||^2 - 2\hat{\boldsymbol{\mu}}_i^{(k)T} \mathbf{W}^{(k+1)} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}) + \text{tr}\left[\tilde{\boldsymbol{\mu}}_i^{(k)} (\mathbf{W}^{(k+1)})^T \mathbf{W}^{(k+1)} \right] \right)^{-1}$$

Maximization of the observed set log-likelihood

The probabilistic PCA model also makes it possible to analytically maximize its likelihood directly and, as a consequence, to express the linear projection of any p-dimensional point onto the d-dimensional subspace in a closed form.

The log-likelihood of the dataset in the model is

$$\begin{aligned}\log p(\mathbf{X}; \mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= \sum_{i=1}^{n} \log p(\bar{\mathbf{x}}_i; \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\ &= -\frac{nD}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma_{\mathbf{x}}| - \frac{1}{2}\sum_{i=1}^{n}(\bar{\mathbf{x}}_i - \boldsymbol{\mu})\Sigma_{\mathbf{x}}^{-1}(\bar{\mathbf{x}}_i - \boldsymbol{\mu})^T\end{aligned}$$

Maximization wrt $\boldsymbol{\mu}$ can be easily done by setting the corresponding gradient to zero, which results into

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

Maximization wrt $\mathbf{W}$ is more complex: however, a closed form solution exists:

$$\mathbf{W}^* = \mathbf{U}_d(\mathbf{L}_d - \sigma^2\mathbf{I})^{1/2} \in \mathbb{R}^{D\times d}$$

where

- $\mathbf{U}_d$ is the $D\times d$ matrix whose columns $1, \dots, d$ are the eigenvectors corresponding to the d largest eigenvalues of the scatter matrix
$$\mathbf{S} \triangleq \frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \in \mathbb{R}^{D\times D}$$
- $\mathbf{L}_d$ is the $d \times d$ diagonal matrix of the d largest eigenvalues $\lambda_1, \dots, \lambda_d$

The columns of $\mathbf{W}^*$ are the eigenvectors $1, \dots, d$, each i scaled by the square root of the difference $\lambda_i - \sigma^2$.

Indeed, any rotation of $\mathbf{W}^*$ in latent space is a solution of the likelihood maximization problem. Hence, the general solution is given by

$$\mathbf{W}^* = \mathbf{U}_d(\mathbf{L}_d - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

where $\mathbf{R}$ is an arbitrary $d \times d$ orthogonal matrix, corresponding to a rotation in $\mathbb{R}^d$.
For what concerns the maximization wrt σ^2, it results

$$\sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i$$

Since eigenvalues provide measures of the dataset variance along the corresponding eigenvector direction, this corresponds to the average variance along the discarded directions.