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 $p(\mathbf{z}|\mathbf{x})$, and this Hypothesis 3 not holding.

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

- a linear projection from the p-dimensional space $\mathbb{R}^p$ 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)

- $\mathbf{z} \in \mathbb{R}^p$ is a latent variable whose distribution is assumed gaussian with 0 mean and unitary covariance matrix: hence $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$
- $\mathbf{W} \in \mathbb{R}^{d \times p}$ is a linear projection of any point in $\mathbb{R}^p$ 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

Consider two random variables $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{z} \in \mathbb{R}^p$ and let

$$\mathbf{y} = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix} \in \mathbb{R}^{d+p}$$

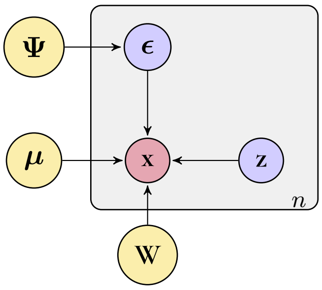

Figure 1: The latent variables $\boldsymbol{\epsilon}$ and $\mathbf{z}$ are normally distributed on the observed and the latent space, respectively: they can be both seen as random noise $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi})$ and $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. 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 $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{z}; \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$.

Assume that $\mathbf{x}$ and $\mathbf{z}$ are jointly multivariate Gaussian; hence, the variable $\mathbf{y}$ has a multivariate Gaussian distribution, i.e., $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}, \Sigma_{\mathbf{y}})$. The mean and covariance of such distribution can be decomposed as:

$$\boldsymbol{\mu}_{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{z}} \end{bmatrix} \in \mathbb{R}^{d+p} \qquad \Sigma_{\mathbf{y}} = \begin{bmatrix} \Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xz}} \\ \Sigma_{\mathbf{zx}} & \Sigma_{\mathbf{z}} \end{bmatrix} \in \mathbb{R}^{(d+p)\times(d+p)},$$

where $\boldsymbol{\mu}_{\mathbf{x}} \in \mathbb{R}^d$, $\boldsymbol{\mu}_{\mathbf{y}} \in \mathbb{R}^p$, $\Sigma_{\mathbf{x}} \in \mathbb{R}^{d\times d}$, $\Sigma_{\mathbf{z}} \in \mathbb{R}^{p\times p}$, $\Sigma_{\mathbf{xz}} \in \mathbb{R}^{d\times p}$, and $\Sigma_{\mathbf{zx}} = \Sigma_{\mathbf{xz}}^T \in \mathbb{R}^{p\times d}$.

It can be shown that the marginal distributions $p(\mathbf{x})$ and $p(\mathbf{z})$ are Gaussian distributions with $E[\mathbf{x}] = \boldsymbol{\mu}_{\mathbf{x}}$ and $E[\mathbf{z}] = \boldsymbol{\mu}_{\mathbf{z}}$. The covariance matrix of the joint distribution can be simplified as:

$$\Sigma_{\mathbf{y}} = \mathop{\mathbb{E}}_{\mathbf{y}}\Big[(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})^T\Big] = \begin{bmatrix} \mathbb{E}_{\mathbf{x}}\Big[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T\Big], \mathbb{E}_{\mathbf{x},\mathbf{z}}\Big[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})^T\Big] \\ \mathbb{E}_{\mathbf{x},\mathbf{z}}\Big[(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T\Big], \mathbb{E}_{\mathbf{z}}\Big[(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})^T\Big] \end{bmatrix}$$

This shows that:

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}}),$$
$$p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \Sigma_{\mathbf{z}}).$$

According to the definition of the multivariate Gaussian distribution, the conditional distribution is also Gaussian, i.e., $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \Sigma_{\mathbf{x}|\mathbf{z}})$ where:

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}} = \boldsymbol{\mu}_{\mathbf{x}} + \Sigma_{\mathbf{xz}}\Sigma_{\mathbf{z}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}) \in \mathbb{R}^d$$
$$\Sigma_{\mathbf{x}|\mathbf{z}} = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xz}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{zx}} = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xz}}\Sigma_{\mathbf{z}}^{-1}\Sigma_{\mathbf{xz}}^T \in \mathbb{R}^{d\times d}$$

and likewise for $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$:

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{z}} + \Sigma_{\mathbf{zx}}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \in \mathbb{R}^p,$$
$$\Sigma_{\mathbf{z}|\mathbf{x}} = \Sigma_{\mathbf{z}} - \Sigma_{\mathbf{zx}}\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{xz}} = \Sigma_{\mathbf{z}} - \Sigma_{\mathbf{xz}}^T\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{xz}} \in \mathbb{R}^{p\times p}.$$

All marginal and conditional distributions turn out to be Gaussian also under the following different hypotheses:

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

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

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

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 $\mathbf{W} \in \mathbb{R}^{d\times p}$,
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

$$\begin{aligned}\Sigma_{\mathbf{z}|\mathbf{x}} &= \left(\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W}\right)^{-1}\\ \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} &= \Sigma_{\mathbf{z}|\mathbf{x}}\left(\mathbf{W}^T\boldsymbol{\Psi}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\end{aligned}$$

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

$$\mathop{\mathbb{E}}_{p(\mathbf{z}|\mathbf{x})}[\mathbf{z}] = \left(\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^T\boldsymbol{\Psi}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

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

Maximization of likelihood in FA

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

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

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

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i$$

However, no closed form solution for 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)

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

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

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

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 $\boldsymbol{\theta} = (\mathrm{W},\boldsymbol{\mu},\boldsymbol{\Psi})$.

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

$$\overline{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i$$

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

$$\begin{aligned}\overline{\mathrm{W}} &= \left(\sum_{i=1}^{n}\overline{\mathrm{x}}_i\boldsymbol{\mu}_{\mathrm{z}_i|\mathrm{x}_i}\right)\left(\sum_{i=1}^{n}\boldsymbol{\mu}_{\mathrm{z}_i\mathrm{z}_i^T|\mathrm{x}_i}\right)^{-1}\\ \overline{\boldsymbol{\Psi}} &= \frac{1}{n}\mathbf{diag}\left(\mathrm{S}-\mathrm{W}\sum_{i=1}^{n}\boldsymbol{\mu}_{\mathrm{z}_i|\mathrm{x}_i}\overline{\mathrm{x}}_i^T\right)\end{aligned}$$

where

1. $\boldsymbol{\mu}_{\mathbf{z}_i|\mathbf{x}_i}$ and $\boldsymbol{\mu}_{\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i}$ are the expectations wrt distribution $p(\mathbf{z}_i|\mathbf{x}_i;\boldsymbol{\theta}^{(k)})$ of $\mathbf{z}_i$ and $\mathbf{z}_i\mathbf{z}_i^T$, respectively

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{z}_i|\mathbf{x}_i} &\triangleq \mathop{\mathbb{E}}_{p(z_i|\mathbf{x}_i;\boldsymbol{\theta}^{(k)})}\left[\, \mathbf{z}_i \,\right] \\
\boldsymbol{\mu}_{\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i} &\triangleq \mathop{\mathbb{E}}_{p(z_i|\mathbf{x}_i;\boldsymbol{\theta}^{(k)})}\left[\, \mathbf{z}_i\mathbf{z}_i^T \,\right]
\end{aligned}$$

2. $\overline{\mathbf{x}}_i$ is the difference between $\mathbf{x}_i$ and the centroid $\overline{\mathbf{x}}$

$$\overline{\mathbf{x}}_i \triangleq \mathbf{x}_i - \overline{\mathbf{x}}$$

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

$$\mathrm{S} \triangleq \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

E-step For the M-step, the conditional expectations $\boldsymbol{\mu}_{\mathbf{z}_i|\mathbf{x}_i}$ and $\boldsymbol{\mu}_{\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i}$ are computed in the E-step. They can be shown to be

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} &= \Sigma_{\mathbf{z}|\mathbf{x}}\Sigma_{\mathbf{z}|\mathbf{x}}\mathrm{W}^T\boldsymbol{\Psi}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \\
\boldsymbol{\mu}_{\mathbf{z}\mathbf{z}^T|\mathbf{x}} &= \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}^T + \Sigma_{\mathbf{z}|\mathbf{x}}
\end{aligned}$$

where, as shown above,

$$\Sigma_{\mathbf{z}|\mathbf{x}} = \left(\mathbf{I} + \mathrm{W}^T\boldsymbol{\Psi}^{-1}\mathrm{W}\right)^{-1}$$

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

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

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,

$$\boldsymbol{\Psi} = \sigma^2\mathbf{I}$$

The resulting model is described graphically in Figure 2.

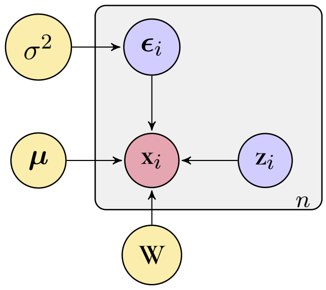

Figure 2: The latent variables $\boldsymbol{\epsilon}$ and $\mathbf{z}$ are normally distributed on the observed and the latent space, respectively: they can be both seen as random noise $p(\boldsymbol{\epsilon};\sigma^2) = \mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\sigma^2\mathbf{I})$ and $p(\mathbf{z})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

Expectation-Maximization for Probabilistic PCA

Clearly, Hypothesis 3 does not hold alsso in this case, and expectation maximization can still be applied to maximize the log-likelihood of the observed dataset $\mathbf{X}$ wrt the parameters 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 $\boldsymbol{\Psi}$.

This results in the following:

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{z}_i|\mathbf{x}_i} &= \beta\Sigma_{\mathbf{z}|\mathbf{x}}\mathrm{W}^T\overline{\mathbf{x}}_i \\
\boldsymbol{\mu}_{\mathbf{z}\mathbf{z}^T|\mathbf{x}} &= \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}^T + \Sigma_{\mathbf{z}|\mathbf{x}}
\end{aligned}$$

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

It can be proved that the algorithm behaves, at each step, as follows (d is the dimensionality, that is the size of data items).

$$\begin{aligned}
&\text{for } i = 1, \dots, n: \\
&\qquad \boldsymbol{\mu}^{(k)}_{\mathbf{z}_i|\mathbf{x}_i} \leftarrow \beta^{(k-1)}\Sigma^{(k-1)}_{\mathbf{z}|\mathbf{x}}\left(\mathrm{W}^{(k-1)}\right)^T\mathbf{x}_i \\
&\qquad \boldsymbol{\mu}^{(k)}_{\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i} \leftarrow \boldsymbol{\mu}^{(k-1)}_{\mathbf{z}_i|\mathbf{x}_i}\boldsymbol{\mu}^{(k-1)^T}_{\mathbf{z}_i|\mathbf{x}_i} + \Sigma^{(k-1)}_{\mathbf{z}|\mathbf{x}} \\
&\mathrm{W}^{(k)} \leftarrow \left(\sum_{i=1}^{n}\mathbf{x}_i\left(\boldsymbol{\mu}^{(k)}_{\mathbf{z}_i|\mathbf{x}_i}\right)^T\right)\left(\sum_{i=1}^{n}\boldsymbol{\mu}^{(k)}_{\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i}\right)^{-1} \\
&\beta^{(k)} \leftarrow nd\left(\sum_{i=1}^{n}\left(\overline{\mathbf{x}}_i\overline{\mathbf{x}}_i^T - 2\boldsymbol{\mu}^{(k)}_{\mathbf{z}_i|\mathbf{x}_i}{}^T\mathrm{W}^{(k)}\overline{\mathbf{x}}_i + \operatorname{tr}\left(\boldsymbol{\mu}^{(k)}_{\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i}\left(\mathrm{W}^{(k)}\right)^T\mathrm{W}^{(k)}\right)\right)\right)^{-1}
\end{aligned}$$

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 D-dimensional point onto the d-dimensional subspace in a closed form.

The log-likelihood of the dataset in the model is

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

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

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

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

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

where

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

The columns of $\mathbf{W}^*$ are the eigenvectors $1, \ldots, d$, 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.