Probabilistic PCA and Factor analysis

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"

Giorgio Gambosi

a.a. 2018-2019

Idea

Introduce a latent variable model to relate a d-dimensional observation vector to a corresponding d'-dimensional gaussian latent variable (with $d' < d$)

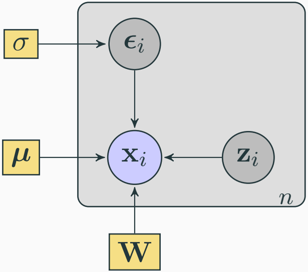
$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where

- $\mathbf{z}$ is a d'-dimensional gaussian latent variable (the “projection” of $\mathbf{x}$ on a lower-dimensional subspace)
- $\mathbf{W}$ is a $d \times d'$ matrix, relating the original space with the lower-dimensional subspace
- $\boldsymbol{\epsilon}$ is a d-dimensional gaussian noise: noise covariance on different dimensions is assumed to be 0. Noise variance is assumed equal on all dimensions: hence $p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$
- $\boldsymbol{\mu}$ is the d-dimensional vector of the means

$\boldsymbol{\epsilon}$ and $\boldsymbol{\mu}$ are assumed independent.

Graphical model



1. $\mathbf{z} \in \mathbb{R}^{d'}, \mathbf{x}, \boldsymbol{\epsilon} \in \mathbb{R}^{d}, d' < d$
2. $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
3. $p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, (isotropic gaussian noise)

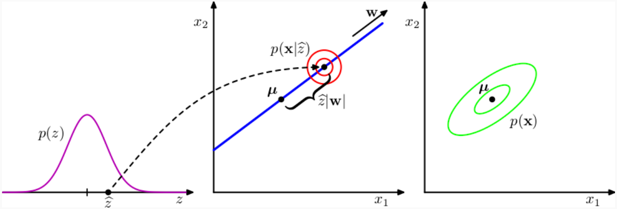
Generative process

This can be interpreted in terms of a generative process

1. sample the latent variable $\mathbf{z} \in \mathbb{R}^{d'}$ from
$$p(\mathbf{z}) = \frac{1}{(2\pi)^{d'/2}} e^{-\frac{||\mathbf{z}||^2}{2}}$$
2. linearly project onto $\mathbb{R}^d$
$$\mathbf{y} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu}$$
3. sample the noise component $\boldsymbol{\epsilon} \in \mathbb{R}^d$ from
$$p(\boldsymbol{\epsilon}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{||\boldsymbol{\epsilon}||^2}{2\sigma^2}}$$
4. add the noise component $\boldsymbol{\epsilon}$
$$\mathbf{x} = \mathbf{y} + \boldsymbol{\epsilon}$$

This results into $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$

Generative process



Let

$$\mathbf{x}_1 \in \mathbb{R}^r \qquad \mathbf{x}_2 \in \mathbb{R}^s \qquad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$$

Assume $\mathbf{x}$ is normally distributed: $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &\in \mathbb{R}^r \\ \boldsymbol{\mu}_2 &\in \mathbb{R}^s \\ \boldsymbol{\Sigma}_{11} &\in \mathbb{R}^{r \times r} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T &\in \mathbb{R}^{r \times s} \\ \boldsymbol{\Sigma}_{22} &\in \mathbb{R}^{s \times s} \end{aligned}$$

Probability recall

Under the above assumptions:

- The marginal distribution $p(\mathbf{x}_1)$ is a gaussian on $\mathbb{R}^r$, with

$$E[\mathbf{x}_1] = \boldsymbol{\mu}_1$$
$$\mathrm{Cov}(\mathbf{x}_1) = \boldsymbol{\Sigma}_{11}$$

- The conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is a gaussian on $\mathbb{R}^r$, with

$$E[\mathbf{x}_1|\mathbf{x}_2] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\mathrm{Cov}(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Probability recall

Under the same hypotheses, the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is a gaussian on $\mathbb{R}^r$, with

$$E[\mathbf{x}_1|\mathbf{x}_2] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

and

$$\mathrm{Cov}(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Latent variable model

The joint distribution is

$$p\left(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix}\right) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{zx}}, \boldsymbol{\Sigma})$$

By definition,

$$\boldsymbol{\mu}_{\mathbf{zx}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{z}} \\ \boldsymbol{\mu}_{\mathbf{x}} \end{bmatrix}$$

- Since $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\boldsymbol{\mu}_{\mathbf{z}} = 0$.
- Since $p(\mathbf{x}) = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$, then

  $$\boldsymbol{\mu}_{\mathbf{x}} = E[\mathbf{x}] = E[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \mathbf{W}E[\mathbf{z}] + \boldsymbol{\mu} + E[\boldsymbol{\epsilon}] = \boldsymbol{\mu}$$

  Hence

  $$\boldsymbol{\mu}_{\mathbf{zx}} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix}$$

For what concerns the distribution covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{zz}} & \boldsymbol{\Sigma}_{\mathbf{zx}} \\ \boldsymbol{\Sigma}_{\mathbf{zx}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{zz}} &= E[(\mathbf{z} - E[\mathbf{z}])(\mathbf{z} - E[\mathbf{z}])^T] = E[\mathbf{z}\mathbf{z}^T] = \mathbf{I} \\
\boldsymbol{\Sigma}_{\mathbf{zx}} &= E[(\mathbf{z} - E[\mathbf{z}])(\mathbf{x} - E[\mathbf{x}])^T] = \mathbf{W}^T \\
\boldsymbol{\Sigma}_{\mathbf{xx}} &= E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}
\end{aligned}$$

Joint distribution

As a consequence, we get

$$\boldsymbol{\mu}_{\mathbf{zx}} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{I} & \mathbf{W}^T \\ \mathbf{W} & \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I} \end{bmatrix}$$

Marginal distribution

The marginal distribution of $\mathbf{x}$ is then $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$

Conditional distribution

The conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ is $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$ with

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$
$$\Sigma_{\mathbf{z}|\mathbf{x}} = \mathbf{I} - \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{W} = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})^{-1}$$

Maximum likelihood for PCA

Setting $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$, the log-likelihood of the dataset in the model is

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

Setting the derivative wrt $\boldsymbol{\mu}$ to zero results into

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

Maximum likelihood for PCA

Maximization wrt $\mathbf{W}$ and σ^2 is more complex: however, a closed form solution exists:

$$\mathbf{W} = \mathbf{U}_{d'}(\mathbf{L}_{d'} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

where

- $\mathbf{U}_{d'}$ is the $d \times d'$ matrix whose columns are the eigenvectors corresponding to the d' largest eigenvalues
- $\mathbf{L}_{d'}$ is the $d' \times d'$ diagonal matrix of the largest eigenvalues
- $\mathbf{R}$ is an arbitrary $d' \times d'$ orthogonal matrix, corresponding to a rotation in the latent space

$\mathbf{R}$ can be interpreted as a rotation matrix in latent space. If $\mathbf{R} = \mathbf{I}$, the columns of $\mathbf{W}$ are the principal components eigenvectors scaled by the variance $\lambda_i - \sigma^2$

For what concerns 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.

The conditional distribution

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu}), \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})^{-1})$$

can be applied. In particular, the conditional expectation

$$E[\mathbf{z}|\mathbf{x}] = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

can be assumed as the latent space point corresponding to $\mathbf{x}$. The projection onto the d'-dimensional subspace can then be performed as

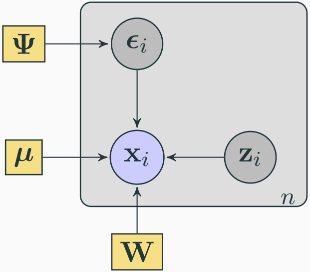
$$\mathbf{x}' = \mathbf{W}E[\mathbf{z}|\mathbf{x}] + \boldsymbol{\mu} = \mathbf{W}\mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \boldsymbol{\mu}$$

Even if the log-likelihood has a closed form maximization, applying the Expectation-Maximization algorithm can be useful in high-dimensional spaces.

Factor analysis

Graphical model

Noise components still gaussian and independent, but with different variance.



1. $\mathbf{z} \in \mathbb{R}^d, \mathbf{x}, \boldsymbol{\epsilon} \in \mathbb{R}^D, d << D$
2. $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
3. $p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, $\boldsymbol{\Psi}$ diagonal (independent gaussian noise)

Generative model

1. sample the vector of factors $\mathbf{z} \in \mathbb{R}^d$ from

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}||\mathbf{z}||^2)$$

2. perform a linear projection onto $\mathbb{R}^D$ (a subspace of dimension d of $\mathbb{R}^D$)

$$\mathbf{y} = \boldsymbol{\Lambda}\mathbf{z} + \boldsymbol{\mu}$$

3. sample the noise component $\boldsymbol{\epsilon} \in \mathbb{R}^D$ from

$$p(\boldsymbol{\epsilon}) = \frac{1}{(2\pi)^{D/2}} \exp(-\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\epsilon})$$

4. add the noise component $\boldsymbol{\epsilon}$

$$\mathbf{x} = \mathbf{y} + \boldsymbol{\epsilon}$$

Model distribution are modified accordingly.

- Joint distribution

$$p\left(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{W} \end{bmatrix}, \begin{bmatrix} \mathbf{I} & \mathbf{W}^T \\ \mathbf{\Lambda} & \mathbf{W}\mathbf{W}^T + \mathbf{\Psi} \end{bmatrix}\right)$$

- Marginal distribution

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

- Conditional distribution The conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ is now $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$ with

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

Maximum likelihood for FA

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

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

Setting the derivative wrt $\boldsymbol{\mu}$ to zero results into

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

Estimating parameters through log-likelihood maximization does not provide a closed form solution for $\mathbf{W}$ and $\boldsymbol{\Psi}$. Iterative techniques such as Expecation-Maximization must be applied.