UNSUPERVISED LEARNING 2011

LECTURE : FACTOR ANALYSIS

Rita Osadchy

Based on Lecture Notes by A. Ng
Motivation

- Distribution comes from MoG
  - Have sufficient amount of data: $m \gg n$
  - Use EM to fit Mixture of Gaussians

- If $m \ll n$
  - difficult to model a single Gaussian
  - much less a mixture of Gaussian
Motivation

- $m$ data points span only a low-dimensional subspace of $\mathbb{R}^n$
- ML estimator of Gaussian parameters:

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x_i$$
$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T$$

- More generally, unless $m$ exceeds $n$ by some reasonable amount, the maximum likelihood estimates of the mean and covariance may be quite poor.

Can’t compute Gaussian Density
Restriction on $\Sigma$

- **Goal:** Fit a reasonable Gaussian model to the data when $m<<n$.

- **Possible solutions:**
  - Limit the number of parameters, assume $\Sigma$ is diagonal.
  - Limit $\Sigma = \sigma^2 I$, where $\sigma^2$ is the parameter under our control.
Contours of a Gaussian Density

General $\Sigma$

Diagonal $\Sigma$

Contours are axis aligned

$\Sigma = \sigma^2 I,$
Correlation in the data

- Restricting $\Sigma$ to be diagonal means modelling the different coordinates of the data as being uncorrelated and independent.

- Often, we would like to capture some interesting correlation structure in the data.
Modeling Correlation

The model we will see today

Correlations

Number of parameters

General $\Sigma$

Factor Analysis

Diagonal $\Sigma$

The model we will see today
Factor Analysis Model

Assume a latent random variable $z \in \mathbb{R}^k$ ($k < n$), $z \sim N(0, I)$

The parameters of the model

$$x \mid z \sim N(\mu + \Lambda z, \Psi)$$

$\mu \in \mathbb{R}^n$ $\Lambda \in \mathbb{R}^{n \times k}$ $\Psi \in \mathbb{R}^{n \times n}$ is diagonal

Equivalently,

$$x = \mu + \Lambda z + \varepsilon$$

$\varepsilon \sim N(0, \Psi)$

$z$ and $\varepsilon$ are independent.
Example of the generative model of $x$

$z \in \mathbb{R}^1$, $x \in \mathbb{R}^2$

$p(z \mid 0,1)$

\[
\begin{bmatrix}
\lambda \\
\mu \\
\psi
\end{bmatrix}
=\begin{bmatrix}
2 \\
1 \\
0
\end{bmatrix},
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
1 & 0 \\
0 & 2
\end{bmatrix}
\]

$x = \mu + \Lambda z + \varepsilon \quad \varepsilon \sim N(0, \Psi)$
Generative process in higher dimensions

- We assume that each data point is generated by sampling a \( k \)-dimension multivariate Gaussian \( z_i \).
- Then, it is mapped to a \( k \)-dimensional affine space of \( \mathbb{R}^n \) by computing \( \mu + \Lambda z_i \).
- Lastly, \( x_i \) is generated by adding covariance \( \Psi \) noise to \( \mu + \Lambda z_i \).
Definitions

- Suppose \( x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \) is r.v., where \( x_1 \in \mathbb{R}^r, x_2 \in \mathbb{R}^s, x \in \mathbb{R}^{r+s} \)

- Suppose \( x \sim N(\mu, \Sigma) \), where

\[
\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}
\]

Here \( \mu_1 \in \mathbb{R}^r, \mu_2 \in \mathbb{R}^s, \Sigma_{11} \in \mathbb{R}^{r\times r}, \Sigma_{12} \in \mathbb{R}^{r\times s}, \ldots \) and \( \Sigma_{12} = \Sigma_{21}^T \)

- Under our assumptions, \( x_1 \) and \( x_2 \) are jointly multivariate Gaussian.
Marginal distribution of $x_1$

$$p(x_1) = \int_{x_2} p(x_1, x_2) dx_2$$

Marginal distributions of Gaussians are themselves Gaussian, hence $x_1 \sim N(\mu_1, \Sigma_{11})$

By definition of the joint covariance of $x_1$ and $x_2$

$$Cov(x) = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E[(x - \mu)(x - \mu)^T] = E \begin{bmatrix} (x_1 - \mu_1) & (x_1 - \mu_1)^T \\ (x_2 - \mu_2) & (x_2 - \mu_2)^T \end{bmatrix}$$

$$= E \begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1)^T & (x_1 - \mu_1)(x_2 - \mu_2)^T \\ (x_2 - \mu_2)(x_1 - \mu_1)^T & (x_2 - \mu_2)(x_2 - \mu_2)^T \end{bmatrix}.$$ 

$$Cov(x_1) = E[(x_1 - \mu_1)(x_1 - \mu_1)^T] = \Sigma_{11}$$
Conditional distribution of $x_1$ given $x_2$

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} \quad \text{where} \quad N(\mu, \Sigma)$$

Referring to the definition of the multivariate Gaussian distribution, it can be shown that

$$x_1 \mid x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2),$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$
Finding the Parameters of FA model

- Assume $z$ and $x$ have a joint Gaussian distribution:

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim N(\mu_{zx}, \Sigma)$$

- We want to find $\mu_{zx}$ and $\Sigma$

$$E[z] = 0 \quad \text{(since } z \sim N(0, I))$$

$$E[x] = E[\mu + \Lambda z + \epsilon] = \mu + \Lambda E[z] + E[\epsilon] = \mu.$$ 

$$E[z] = \begin{bmatrix} \bar{0} \\ \mu \end{bmatrix}$$

$$E[z] = \begin{bmatrix} \bar{0} \\ \mu \end{bmatrix}$$

$$\mu_{zx} = E\begin{bmatrix} z \\ x \end{bmatrix} = \begin{bmatrix} \bar{0} \\ \mu \end{bmatrix}$$
Finding $\Sigma$

- We need to calculate
  - upper left block
    \[
    \Sigma_{zz} = E[(z - E[z])(z - E[z])^T]
    \]
  - upper-right block
    \[
    \Sigma_{zx} = E[(z - E[z])(x - E[x])^T]
    \]
  - lower-right block
    \[
    \Sigma_{xx} = E[(x - E[x])(x - E[x])^T]
    \]

$z \sim N(0, I)$

\[
\Sigma_{zz} = \text{Cov}(z) = I
\]
Finding $\Sigma_{zx}$

\[ E[(z - E[z])(x - E[x])^T] = E[z(\mu + \Lambda z + \varepsilon - \mu)^T] \]

\[ = E[zz^T] \Lambda + E[z\varepsilon^T] \]

$\text{Cov}(z)$

\[ \text{independent} \]

\[ E[z]E[\varepsilon] = 0 \]

\[ = \Lambda^T \]
Finding $\Sigma_{xx}$

Similarly,

$$
\Sigma_{xx} = E[(x - E[x])(x - E[x])^T]
= E[(\mu + \Lambda z + \varepsilon - \mu)(\mu + \Lambda z + \varepsilon - \mu)^T]
= E[\Lambda z z^T \Lambda^T + \varepsilon \varepsilon^T \Lambda^T - \Lambda z \varepsilon^T + \varepsilon \varepsilon^T]
= \Lambda E[z z^T] \Lambda^T + E[\varepsilon \varepsilon^T] = \Lambda \Lambda^T + \Psi
$$
Finding the parameters (cont.)

Putting everything together, we have that,

\[
\begin{bmatrix}
  z \\
  x
\end{bmatrix} \sim N\left(\begin{bmatrix}
  \tilde{0} \\
  \mu
\end{bmatrix}, \begin{bmatrix}
  I & \Lambda^T \\
  \Lambda & \Lambda\Lambda^T + \Psi
\end{bmatrix}\right)
\]

We also see that the marginal distribution of \( x \) is given by

\[x \sim N(\mu, \Lambda\Lambda^T + \Psi)\]

Thus, given a training set \( \{x_i\}_{i=1}^m \) log likelihood of the parameters is:

\[
l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^m \frac{1}{(2\pi)^{n/2} |\Lambda\Lambda^T + \Psi|} \exp\left(-\frac{1}{2} (x_i - \mu)^T (\Lambda\Lambda^T + \Psi)(x_i - \mu)\right)
\]
Finding the parameters (cont.)

To perform maximum likelihood estimation, we would like to maximize this quantity with respect to the parameters.

But maximizing this formula explicitly is hard, and we are aware of no algorithm that does so in closed-form.

So, we will instead use the EM algorithm.
EM for Factor Analysis

- **E-step:**

\[ Q_i(z_i) = p(z_i | x_i, \theta) \]

- **M-step:**

\[ \theta = \arg \max_{\theta} \sum_i \int Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)} dz_i \]
E-step (EM for FA)

- We need to compute $Q_i(z_i) = p(z_i | x_i; \mu, \Lambda, \Psi)$
- Using a conditional distribution of a Gaussian we find that $z_i | x_i \sim N(\mu_{z_i|x_i}, \Sigma_{z_i|x_i})$

$$
\begin{align*}
\mu_{12} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \\
\Sigma_{12} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
\Sigma &= \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}
\end{align*}
$$

$$
\begin{align*}
\mu_{z_i|x_i} &= \mu - \Lambda^T (\Lambda^T \Lambda + \Psi)^{-1}(x_i - \mu) \\
\Sigma_{z_i|x_i} &= I - \Lambda^T (\Lambda^T \Lambda + \Psi)^{-1}\Lambda
\end{align*}
$$

$$
Q_i(z_i) = \frac{1}{(2\pi)^{2k}\left|\Sigma_{z_i|x_i}\right|^{1/2}} \exp\left(-\frac{1}{2}(z_i - \mu_{z_i|x_i})^T \Sigma_{z_i|x_i}^{-1}(z_i - \mu_{z_i|x_i})\right)
$$
M-step (EM for FA)

Maximize:

\[
\sum_{i=1}^{m} \int_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \mu, \Lambda, \Psi)}{Q_i(z_i)} \, dz_i
\]

with respect to the parameters \( \mu, \Lambda, \Psi \)

- We will work out the optimization with respect to \( \Lambda \)

- Derivations of the updates for \( \mu, \Psi \) is an exercise (Do it!)
Update for $\Lambda$

$$
\sum_{i=1}^{m} \int_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \mu, \Lambda, \Psi)}{Q_i(z_i)} \, dz_i
$$

$$
= \sum_{i=1}^{m} \int_{z_i} Q_i(z_i) \left[ \log p(x_i \mid z_i; \mu, \Lambda, \Psi) + \log p(z_i) - \log Q_i(z_i) \right] \, dz_i
$$

Expectation with respect to $z_i$, drawn from $Q_i$

$$
= \sum_{i=1}^{m} \mathbb{E}_{z_i \sim Q_i} \left[ \log p(x_i \mid z_i; \mu, \Lambda, \Psi) + \log p(z_i) - \log Q_i(z_i) \right]
$$
Update for $\Lambda$ (cont.)

$$\sum_{i=1}^{m} E_{z_i \sim Q_i} [\log p(x_i \mid z_i ; \mu, \Lambda, \Psi) + \log p(z_i) - \log Q_i(z_i)]$$

Remember that We want to maximize this expression with respect to $\Lambda$

$$\sum_{i=1}^{m} E_{z_i \sim Q_i} [\log p(x_i \mid z_i ; \mu, \Lambda, \Psi)]$$

$$= \sum_{i=1}^{m} E \left[ \log \frac{1}{(2\pi)^{n/2} |\Psi|^{1/2}} \exp \left( -\frac{1}{2} (x_i - \mu - \Lambda z_i)^T \Psi^{-1} (x_i - \mu - \Lambda z_i) \right) \right]$$

$$= \sum_{i=1}^{m} E \left[ -\frac{1}{2} \log |\Psi| - \frac{n}{2} \log(2\pi) - \frac{1}{2} (x_i - \mu - \Lambda z_i)^T \Psi^{-1} (x_i - \mu - \Lambda z_i) \right]$$

Do not depend on $\Lambda$
Update for Λ (cont.)

Take derivative with respect to Λ

$\nabla_\Lambda \sum_{i=1}^{m} - \mathbb{E} \left[ \frac{1}{2} (x_i - \mu - \Lambda z_i)^T \Psi^{-1} (x_i - \mu - \Lambda z_i) \right]$

Simplify:

$\sum_{i=1}^{m} \nabla_\Lambda \mathbb{E} \left[ - \operatorname{tr} \frac{1}{2} z_i^T \Lambda^T \Psi^{-1} \Lambda z_i + \operatorname{tr} z_i^T \Lambda^T \Psi^{-1} (x_i - \mu) \right]$

$\operatorname{tr} a = a, \ a \in \mathbb{R}$;
Update for $\Lambda$ (cont.)

$$\sum_{i=1}^{m} \nabla_{\Lambda} E \left[ -\operatorname{tr} \frac{1}{2} z_i^T \Lambda^T \Psi^{-1} \Lambda z_i + \operatorname{tr} z_i^T \Lambda^T \Psi^{-1} (x_i - \mu) \right]$$

$$= \sum_{i=1}^{m} \nabla_{\Lambda} E \left[ -\operatorname{tr} \frac{1}{2} \Lambda^T \Psi^{-1} \Lambda z_i z_i^T + \operatorname{tr} \Lambda^T \Psi^{-1} (x_i - \mu) z_i^T \right]$$

$$= \sum_{i=1}^{m} E \left[ -\Psi^{-1} \Lambda z_i z_i^T + \Psi^{-1} (x_i - \mu) z_i^T \right]$$
Update for $\Lambda$ (cont.)

$$\sum_{i=1}^{m} E\left[-\Psi^{-1}\Lambda z_{i} z_{i}^{T} + \Psi^{-1}(x_{i} - \mu)z_{i}^{T}\right]$$

Setting this to zero and simplifying, we get:

$$\sum_{i=1}^{m} \Lambda E_{z_{i} \sim Q_{i}} \left[z_{i} z_{i}^{T}\right] = \sum_{i=1}^{m} (x_{i} - \mu) E_{z_{i} \sim Q_{i}} \left[z_{i}^{T}\right]$$

Solving for $\Lambda$, we obtain:

$$\Lambda = \left(\sum_{i=1}^{m} (x_{i} - \mu) E_{z_{i} \sim Q_{i}} \left[z_{i}^{T}\right]\right)\left(\sum_{i=1}^{m} E_{z_{i} \sim Q_{i}} \left[z_{i} z_{i}^{T}\right]\right)^{-1}$$

Since $Q$ is Gaussian with mean $\mu_{z_{i}|x_{i}}$ and covariance $\Sigma_{z_{i}|x_{i}}$

$$E_{z_{i} \sim Q_{i}} \left[z_{i}^{T}\right] = \mu_{z_{i}|x_{i}}^{T}$$

$$E_{z_{i} \sim Q_{i}} \left[z_{i} z_{i}^{T}\right] = \mu_{z_{i}|x_{i}} \mu_{z_{i}|x_{i}}^{T} + \Sigma_{z_{i}|x_{i}}$$

hence,

$$\text{Cov}(Y) = E[YY^{T}] - E[Y]E[Y^{T}]$$

$$E[YY^{T}] = E[Y]E[Y^{T}] + \text{Cov}(Y)$$
Update for $\Lambda$ (cont.)

$$E_{z_i \sim Q_i} [z_i z_i^T] = \mu_{z_i|x_i} \mu_{z_i|x_i}^T + \Sigma_{z_i|x_i}$$

$$E_{z_i \sim Q_i} [z_i^T] = \mu_{z_i|x_i}^T$$

\[
\Lambda = \left( \sum_{i=1}^{m} (x_i - \mu) E_{z_i \sim Q_i} [z_i^T] \right) \left( \sum_{i=1}^{m} E_{z_i \sim Q_i} [z_i z_i^T] \right)^{-1}
\]

\[
\Lambda = \left( \sum_{i=1}^{m} (x_i - \mu) \mu_{z_i|x_i}^T \right) \left( \sum_{i=1}^{m} \mu_{z_i|x_i} \mu_{z_i|x_i}^T + \Sigma_{z_i|x_i} \right)^{-1}
\]
M-step updates for $\mu$ and $\Psi$

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x_i$$

Doesn’t depend on $Q_i(z_i) = p(z_i | x_i; \mu, \Lambda, \Psi)$, hence can be computed once for all the iterations.

$$\Phi = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T - x_i \mu_{z_i|x_i}^T \Lambda^T - \Lambda \mu_{z_i|x_i} x_i^T + \Lambda(\mu_{z_i|x_i} \mu_{z_i|x_i}^T + \Sigma_{z_i|x_i}) \Lambda^T$$

The diagonal

$$\Psi_{ii} = \Phi_{ii}$$

(contains only diagonal entries)
Probabilistic PCA

- Probabilistic, generative view of data

\[ p(z) = \mathcal{N}(z|0, I) \]

\[ p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 I) \]

\[ x \in \mathbb{R}^D, \ z \in \mathbb{R}^M \]

\[ x = Wz + \mu + \epsilon \]
Compare

- Probabilistic PCA
  
  \[ p(z) = \mathcal{N}(z|0, I) \]
  
  \[ p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 I) \]

- FA
  
  \[ p(z) = \mathcal{N}(z|0, I) \]
  
  \[ p(x|z) = \mathcal{N}(x|Wz + \mu, \Psi) \]

spherical

diagonal, axis-aligned
Probabilistic PCA

- The columns of $W$ are the principle components.

- Can be found using
  - ML in closed form
  - EM (more efficient when only few eigenvectors are required, avoids evaluation of data covariance matrix)
  - Other advantages (see Bishop, Ch. 12.2)