Lecture 8: Principle Component Analysis and Factor Analysis

Feng Li

Shandong University

fli@sdu.edu.cn

January 7, 2021
Outline

1. Dimensionality Reduction
2. Principle Component Analysis
3. Conditional Gaussian and Marginal Gaussian
4. Factor Analysis
5. EM Algorithm for Factor Analysis
Dimensionality Reduction

- Usually considered an unsupervised learning method
- Used for learning the low-dimensional structures in the data

- Also useful for “feature learning” or “representation learning” (learning a better, often smaller-dimensional, representation of the data), e.g.,
  - Documents using topic vectors instead of bag-of-words vectors
  - Images using their constituent parts (faces - eigenfaces)
- Can be used for speeding up learning algorithms
Dimensionality Reduction (Contd.)

- Exponentially large # of examples required to “fill up” high-dim spaces

- Fewer dimensions ⇒ Less chances of overfitting ⇒ Better generalization

- Dimensionality reduction is a way to beat the curse of dimensionality
A projection matrix $U = [u_1 u_2 \cdots u_K]$ of size $D \times K$ defines $K$ linear projection direction.

Use $U$ to transform $x^{(i)} \in \mathbb{R}^D$ into $z^{(i)} \in \mathbb{R}^K$.

$$z^{(i)} = U^T x^{(i)} = [u_1^T x^{(i)}, u_2^T x^{(i)}, \cdots, u_K^T x^{(i)}]^T$$

is a $K$-dim projection of $x^{(i)}$.

$z^{(i)} \in \mathbb{R}^K$ is also called low-dimensional “embedding” of $x^{(i)} \in \mathbb{R}^D$. 
Linear Dimensionality Reduction

- $X = [x^{(1)} \ x^{(2)} \ldots \ x^{(N)}]$ is $D \times N$ matrix denoting all the $N$ data points.
- $Z = [z^{(1)} \ z^{(2)} \ldots \ z^{(N)}]$ is $K \times N$ matrix denoting embeddings of the data points.
- With this notation, the figure on previous slide can be re-drawn as:

  \[
  Z = U^T \times X
  \]

- How do we learn the “best” projection matrix $U$?
- What criteria should we optimize for when learning $U$?
- Principle Component Analysis (PCA) is an algorithm for doing this.
Principle Component Analysis (PCA)

- PCA is a technique widely used for applications such as dimensionality reduction, lossy data compression, feature extraction, and data visualization.
- Two commonly used definitions:
  - Learning projection directions that capture maximum variance in data
  - Learning projection directions that result in smallest reconstruction error.
- Can also be seen as changing the basis in which the data is represented (and transforming the features such that new features become decorrelated).
Consider \( x^{(i)} \in \mathbb{R}^D \) on a one-dim subspace defined by \( u_1 \in \mathbb{R}^D \) (\( \| u_1 \| = 1 \)).

Projection of \( x^{(i)} \) along a one-dim subspace

\[
\frac{1}{N} \sum_{i=1}^{N} u_1^T x^{(i)} = u_1^T \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = u_1^T \mu
\]
Variance Captured by Projections

- Variance of the projected data

\[ \frac{1}{N} \sum_{i=1}^{N} (u_1^T x^{(i)} - u_1^T \mu)^2 = \frac{1}{N} \sum_{i=1}^{N} [u_1^T (x^{(i)} - \mu)]^2 = u_1^T S u_1 \]

- \( S \) is the \( D \times D \) data covariance matrix

\[ S = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \]

- Variance of the projected data ("spread" of the yellow points)
- If data already centered at \( \mu = 0 \), then \( S = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}(x^{(i)})^T \)
Optimization Problem

- We want $u_1$ s.t. the variance of the projected data is maximized

$$\max_{u_1} u_1^T Su_1$$
$$s.t. \quad u_1^T u_1 = 1$$

- The method of Lagrange multipliers

$$L(u_1, \lambda_1) = u_1^T Su_1 - \lambda_1 (u_1^T u_1 - 1)$$

where $\lambda_1$ is a Lagrange multiplier
Taking the derivative w.r.t. $u_1$ and setting to zero gives

$$Su_1 = \lambda_1 u_1$$

Thus $u_1$ is an eigenvector of $S$ (with corresponding eigenvalue $\lambda_1$)

But which of $S$’s eigenvectors it is?

Note that since $u_1^T u_1 = 1$, the variance of projected data is

$$u_1^T Su_1 = \lambda_1$$

Var. is maximized when $u_1$ is the top eigenvector with largest eigenvalue

The top eigenvector $u_1$ is also known as the first Principle Component (PC)

Other directions can also be found likewise (with each being orthogonal to all previous ones) using the eigendecomposition of $S$ (this is PCA)
Steps in Principle Component Analysis

- Center the data (subtract the mean \( \mu = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \) from each data point)
- Compute the covariance matrix

\[
S = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} x^{(i)^T} = \frac{1}{N} XX^T
\]

- Do an eigendecomposition of the covariance matrix \( S \)
- Take first \( K \) leading eigenvectors \( \{ u_{l} \}_{l=1,\ldots,K} \) with eigenvalues \( \{ \lambda_{l} \}_{l=1,\ldots,K} \)
- The final \( K \) dim. projection of data is given by

\[
Z = U^T X
\]

where \( U \) is \( D \times K \) and \( Z \) is \( K \times N \)
PCA as Minimizing the Reconstruction Error

- Assume complete orthonormal basis vector $u_1, u_2, \cdots, u_D$, each $u_l \in \mathbb{R}^D$

- We can represent each data point $x^{(i)} \in \mathbb{R}^D$ exactly using the new basis

$$x^{(i)} = \sum_{l=1}^{D} z_l^{(i)} u_l$$

$$\begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_D^{(i)} \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & \cdots & u_D \end{bmatrix} \begin{bmatrix} z_1^{(i)} \\ z_2^{(i)} \\ \vdots \\ z_D^{(i)} \end{bmatrix}$$

- Denoting $z^{(i)} = [z_1^{(i)} \cdots z_D^{(i)}]^T$, $U = [u_1 \cdots u_D]$, and using $U^T U = I$

$$x^{(i)} = U z^{(i)} \text{ and } z^{(i)} = U^T x^{(i)}$$

- Also note that each component of vector $z^{(i)}$ is $z_l^{(i)} = u_l^T x^{(i)}$
Reconstruction of Data from Projections

- Reconstruction of $x^{(i)}$ from $z^{(i)}$ will be exact if we use all $D$ basis vectors.
- Will be approximate if we only use $K < D$ basis vectors:
  \[ x^{(i)} \approx \sum_{l=1}^{K} z_l^{(i)} u_l \]

- Let's use $K = 1$ basis vector. Then, the one-dim embedding of $x^{(i)}$ is
  \[ z^{(i)} = u_1^T x^{(i)} \quad (z^{(i)} \in \mathbb{R}) \]

- We can now try to “reconstruct” $x^{(i)}$ from its embedding $z^{(i)}$ as follows
  \[ \tilde{x}^{(i)} = u_1 z^{(i)} = u_1 u_1^T x^{(i)} \]

- Total error or “loss” in reconstructing all the data points
  \[ \ell(u_1) = \sum_{i=1}^{N} \| x^{(i)} - \tilde{x}^{(i)} \|^2 = \sum_{i=1}^{N} \| x^{(i)} - u_1 u_1^T x^{(i)} \|^2 \]
We want to find $u_1$ that minimize the reconstruction error

$$\ell(u_1) = \sum_{i=1}^{N} \| x^{(i)} - u_1 u_1^T x^{(i)} \|^2 = \sum_{i=1}^{N} \left( -u_1^T x^{(i)} (x^{(i)})^T u_1 + (x^{(i)})^T x^{(i)} \right)$$

by using $u_1^T u_1 = 1$

Minimizing the error of reconstructing all the data points is equivalent to

$$\max_{u_1: \|u_1\|^2=1} u_1^T \left( \sum_{n=1}^{N} x^{(i)} (x^{(i)})^T \right) u_1 = \max_{u_1: \|u_1\|^2=1} u_1^T S u_1$$

where $S$ is the covariance matrix of the data (which are assumed to be centered)

It is the same objective that we had when we maximized the variance
Revisiting Gaussian

- Gaussian distribution with a single variable

\[ \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{1}{2\sigma^2}(x - \mu)^2 \right) \]

where \( \mu \) is the mean and \( \sigma^2 \) is the variance

- \( n \)-dimensional multivariate Gaussian distribution

\[ \mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right) \]

where \( \mu \) is the \( n \)-dimensional mean vector and \( \Sigma \) is the \( n \times n \)-dimensional covariance matrix
Central limit theorem

Subject to certain mild conditions, the sum of a set of random variables has a distribution increasingly approaching Gaussian as the number of the variables increases.

Figure: Consider $N$ random variables $x_1, x_2, \cdots, x_N$ each of which has a uniform distribution over $[0, 1]$. The distribution of their mean $\frac{1}{N} \sum_{i=1}^{N} x_i$ tends to a Gaussian as $N \to \infty$. 
The following Gaussian integrals have closed-form solutions

\[
\int_{\mathbb{R}^n} \mathcal{N}(x; \mu, \Sigma) \, dx = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathcal{N}(x; \mu, \Sigma) \, dx_1 \cdots dx_n = 1
\]

\[
\int_{\mathbb{R}^n} x_i \mathcal{N}(x; \mu, \Sigma) \, dx = \mu_i, \quad \forall i = 1, 2, \cdots, n
\]

\[
\int_{\mathbb{R}^n} (x_i - \mu_i)(x_j - \mu_j) \mathcal{N}(x; \mu, \Sigma) \, dx = \Sigma_{ij}
\]
Revisiting Gaussian (Contd.)

- The functional dependence of the Gaussian on $x$ is through the quadratic form

$$\Delta^2 = (x - \mu)^T \Sigma (x - \mu)$$

where $\Delta$ is called the Mahalanobis distance from $x$ to $\mu$

- $\Sigma$ is symmetric such that
  - All eigenvalues of $\Sigma$, i.e., $\lambda_1, \lambda_2, \cdots, \lambda_D$, are real
  - Eigenvectors (i.e., $u_1, u_2, u_D$) corresponding to distinct eigenvalues are orthogonal
An important property

- If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian

\[
\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)
\]
\[
\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}
\]

- Similarly, the marginal distribution of either set is also Gaussian

\[
\mathbb{E}[x_a] = \mu_a
\]
\[
\text{cov}[x_a] = \Sigma_{aa}
\]
Conditional Gaussian Distribution

- $x \sim \mathcal{N}(\mu, \Sigma)$
- Partition $x$ into two disjoint subsets $x_a$ and $x_b$

\[
x = \begin{bmatrix} x_a \\ x_b \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}
\]

- Precision matrix

\[
\Lambda := \Sigma^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}
\]

where $\Lambda_{ab} = \Lambda_{ba}$
Conditional Gaussian Distribution (Contd.)

- $n$-dimensional multivariate Gaussian distribution

$$
\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
$$

where $\mu$ is the $n$-dimensional mean vector and $\Sigma$ is the $n \times n$-dimensional covariance matrix.

- If the conditional probability of $x_a$ conditioned on $x_b$ is a Gaussian

$$
\mathcal{N}(x_a \mid x_b; \mu_{a\mid b}, \Sigma_{a\mid b}) = \frac{1}{(2\pi)^{\frac{n_a}{2}} |\Sigma_{a\mid b}|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (x - \mu_{a\mid b})^T \Sigma_{a\mid b}^{-1} (x - \mu_{a\mid b}) \right)
$$

where $\mu_{a\mid b}$ is the $n_a$-dimensional conditional mean vector of $x_a$ and $\Sigma_{a\mid b}$ is the $n_a \times n_a$-dimensional conditional covariance matrix.
A quadratic form of $x_a$

$$-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)$$

$$= -\frac{1}{2} \left( \begin{bmatrix} x_a \\ x_b \end{bmatrix} - \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \right)^T \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \left( \begin{bmatrix} x_a \\ x_b \end{bmatrix} - \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \right)$$

$$= -\frac{1}{2} (x_a - \mu_a)^T \Lambda_{aa} (x_a - \mu_a) - (x_a - \mu_a)^T \Lambda_{ab} (x_b - \mu_b)$$

$$- \frac{1}{2} (x_b - \mu_b)^T \Lambda_{bb} (x_b - \mu_b)$$

$$= -\frac{1}{2} x_a^T \Lambda_{aa} x_a + x_a^T (\Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b)) + \text{const}$$
A quadratic form of $x_a$

$$-rac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)$$

$$= - \frac{1}{2}x_a^T \Lambda_{aa} x_a + x_a^T (\Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b)) + \text{const}$$

Referring to

$$-rac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) = - \frac{1}{2}x^T \Sigma^{-1}x + x^T \Sigma^{-1} \mu + \text{const}$$

The covariance of $p(x_a \mid x_b)$ is given by

$$\Sigma_{a\mid b} = \Lambda_{aa}^{-1}$$
Conditional Gaussian Distribution (Contd.)

- A quadratic form of $x_a$

$$-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)$$

$$= -\frac{1}{2} x_a^T \Lambda_{aa} x_a + x_a^T (\Lambda_{aa} \mu_a - \Lambda_{ab}(x_b - \mu_b)) + \text{const}$$

- Referring to

$$-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) = -\frac{1}{2} x^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu + \text{const}$$

- The mean of $p(x_a \mid x_b)$ is given by

$$\mu_{a \mid b} = \Sigma_{a \mid b}(\Lambda_{aa} \mu_a - \Lambda_{ab}(x_b - \mu_b)) = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_b)$$
Since

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}^{-1} = \begin{bmatrix}
M & -MBD^{-1} \\
-D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1}
\end{bmatrix}
\]

where \( M = (A - BD^{-1}C)^{-1} \) is known as the Schur complement.

Then

\[
\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}
\]

\[
\Lambda_{ab} = -\left( \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \right)^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}
\]

All in all,

\[
\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)
\]

\[
\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}
\]

Check the normalization item by yourselves.
Marginal Gaussian Distribution

- \(n\)-dimensional multivariate Gaussian distribution

\[
\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right)
\]

where \(\mu\) is the \(n\)-dimensional mean vector and \(\Sigma\) is the \(n \times n\)-dimensional covariance matrix

- Marginal Gaussian

\[
p(x_a) = \int_{\mathbb{R}^{n_b}} p(x_a, x_b) dx_b
\]

- If the marginal probability of \(x_a\) is a Gaussian

\[
\mathcal{N}(x_a; \bar{\mu}_a, \Sigma_a) = \frac{1}{(2\pi)^{n/2}|\Sigma_a|^{1/2}} \exp \left( -\frac{1}{2}(x - \bar{\mu}_a)^T \Sigma_a^{-1}(x - \bar{\mu}_a) \right)
\]
Marginal Gaussian Distribution (Contd.)

- Recalling the quadratic form of $x_a$

$$-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) = -\frac{1}{2} (x_a - \mu_a)^T \Lambda_{aa} (x_a - \mu_a)$$
$$-(x_a - \mu_a)^T \Lambda_{ab} (x_b - \mu_b)$$
$$-\frac{1}{2} (x_b - \mu_b)^T \Lambda_{bb} (x_b - \mu_b)$$

- Picking out all items involving $x_b$

$$-\frac{1}{2} x_b^T \Lambda_{bb} x_b + x_b^T m = -\frac{1}{2} (x_b - \Lambda_{bb}^{-1} m)^T \Lambda_{bb} (x_b - \Lambda_{bb}^{-1} m) + \frac{1}{2} m^T \Lambda_{bb}^{-1} m$$

where $m = \Lambda_{bb} \mu_b - \Lambda_{ba} (x_a - \mu_a)$. 
Taking the exponential of this quadratic form, the integration over $x_b$ can be defined as

$$\int \exp \left( -\frac{1}{2} (x_b - \Lambda_{bb}^{-1} m)^T \Lambda_{bb} (x_b - \Lambda_{bb}^{-1} m) \right) dx_b$$

It is the integral over an unnormalized Gaussian, and hence the result will be the reciprocal of the normalization coefficient which depends only on the determinant of the covariance matrix.
Combining $\frac{1}{2} m^T \Lambda_{bb}^{-1} m$ with the remaining terms depending on $x_a$

$$\frac{1}{2} [\Lambda_{bb} \mu_b - \Lambda_{ba} (x_a - \mu_a)]^T \Lambda_{bb}^{-1} [\Lambda_{bb} \mu_b - \Lambda_{ba} (x_a - \mu_a)]$$

$$- \frac{1}{2} x_a^T \Lambda_{aa} x_a + x_a^T (\Lambda_{aa} \mu_a + \Lambda_{ab} \mu_b) + \text{const}$$

$$= - \frac{1}{2} x_a^T (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba}) x_a$$

$$+ x_a^T (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba}) \mu_a + \text{const}$$

Therefore

$$\Sigma_a = (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba})^{-1} = \Sigma_{aa}$$

$$\bar{\mu}_a = \Sigma_a(\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba}) \mu_a$$
Factor Analysis Model

- \( x = \mu + \Lambda z + \varepsilon \)
  - \( x \in \mathbb{R}^n, \mu \in \mathbb{R}^n, \Lambda \in \mathbb{R}^{n \times k}, z \in \mathbb{R}^k, \varepsilon \in \mathbb{R}^n \)
  - \( \Lambda \) is the factor loading matrix
  - \( z \sim \mathcal{N}(0, I) \) (zero-mean independent normals, with unit variance)
  - \( \varepsilon \sim \mathcal{N}(0, \Psi) \) where \( \Psi \) is a diagonal matrix (the observed variables are independent given the factors)

How do we get the training data \( \{x^{(i)}\}_i \)?
- Generate \( \{z^{(i)}\}_i \) according to a multivariate Gaussian distribution \( \mathcal{N}(0, I) \)
- Map \( \{z^{(i)}\}_i \) into a \( n \)-dimensional affine space by \( \Lambda \) and \( \mu \)
- Generate \( \{x^{(i)}\}_i \) by sampling the above affine space with noise \( \varepsilon \)

Equivalently,

\[
\begin{align*}
z & \sim \mathcal{N}(0, I) \\
x | z & \sim \mathcal{N}(\mu + \Lambda z, \Psi)
\end{align*}
\]
Higher Dimension But Less Data

- Consider a case with $n \gg m$
  - The given training data span only a low-dimensional subspace of $\mathbb{R}^n$
- If we model the data as Gaussian and estimate the mean and covariance using MLE

$$
\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}
$$

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

we may observe that $\Sigma$ may be singular such that $\Sigma^{-1}$ does not exist and $1/|\Sigma|^{1/2} = 1/0$

$$
p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
$$
z and x have a joint Gaussian distribution

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

Question: How to calculate \( \mu_{zx} \) and \( \Sigma \)?

Since \( E[z] = 0 \), we have

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

and then

\[
\mu_{zx} = \begin{bmatrix} 0 \\ \mu \end{bmatrix}
\]
Since \( z \sim \mathcal{N}(0, I) \), \( \mathbb{E}[zz^T] = \text{Cov}(z) \), and \( \mathbb{E}[z\epsilon^T] = \mathbb{E}[z]\mathbb{E}[\epsilon^T] = 0 \),

\[
\Sigma_{zz} = \mathbb{E}[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^T] = \text{Cov}(z) = I
\]

\[
\Sigma_{xx} = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]
\]

\[
= \mathbb{E}[(\mu + \Lambda z + \epsilon - \mu)(\mu + \Lambda z + \epsilon - \mu)^T]
\]

\[
= \mathbb{E}[\Lambda zz^T\Lambda^T + \epsilon z^T\Lambda^T + \Lambda z\epsilon^T + \epsilon\epsilon^T]
\]

\[
= \Lambda \mathbb{E}[zz^T]\Lambda^T + \mathbb{E}[\epsilon\epsilon^T]
\]

\[
= \Lambda \Lambda^T + \Psi
\]

\[
\Sigma_{zx} = \mathbb{E}[(z - \mathbb{E}[z])(x - \mathbb{E}[x])^T]
\]

\[
= \mathbb{E}[z(\mu + \Lambda z + \epsilon - \mu)^T]
\]

\[
= \mathbb{E}[zz^T]\Lambda^T + \mathbb{E}[z\epsilon^T]
\]

\[
= \Lambda^T
\]
Putting everything together, we therefore have

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

Then, \( x \sim \mathcal{N}(\mu, \Lambda \Lambda^T + \Psi) \)

Log-likelihood function

\[
\ell(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2}|\Sigma_{xx}|^{1/2}} \exp \left( -\frac{1}{2} (x^{(i)} - \mu)^T \Sigma_{xx}^{-1} (x^{(i)} - \mu) \right)
\]
Repeat the following step until convergence

- (E-step) For each $i$, set
  \[ Q_i(z^{(i)}) := p(z^{(i)} \mid x^{(i)}; \theta) \]

- (M-step) set
  \[ \theta := \arg \max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \]
Recall that if

\[
\begin{bmatrix}
    x_a \\
    x_b
\end{bmatrix} \sim \mathcal{N}
    \left(
        \mu = \begin{bmatrix}
            \mu_a \\
            \mu_b
        \end{bmatrix},
        \Sigma = \begin{bmatrix}
            \Sigma_{aa} & \Sigma_{ab} \\
            \Sigma_{ba} & \Sigma_{bb}
        \end{bmatrix}
    \right)
\]

we then have

\[
x_a | x_b \sim \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})
\]

where

\[
\mu_{a|b} = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b)
\]
\[
\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}
\]
Since

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

we have

\[z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi \sim \mathcal{N}\left( \mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}} \right)\]

where

\[
\begin{align*}
\mu_{z^{(i)}|x^{(i)}} &= \Lambda^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) \\
\Sigma_{z^{(i)}|x^{(i)}} &= I - \Lambda^T (\Lambda\Lambda^T + \Psi)^{-1} \Lambda
\end{align*}
\]

• Calculate \(Q_i(z^{(i)})\) in the E-step

\[
Q_i(z^{(i)}) = \frac{\exp\left( -\frac{1}{2}(z^{(i)} - \mu_{z^{(i)}|x^{(i)}})^T \Sigma_{z^{(i)}|x^{(i)}}^{-1} (z^{(i)} - \mu_{z^{(i)}|x^{(i)}}) \right)}{(2\pi)^{n/2} |\Sigma_{z^{(i)}|x^{(i)}}|^{1/2}}
\]
In M-step, we maximize the following equation with respect to \( \mu, \Lambda, \) and \( \Psi \)

\[
\sum_{i=1}^{m} \int_{z(i)} Q_i(z(i)) \log \frac{p(x(i), z(i); \mu, \Lambda, \Psi)}{Q_i(z(i))} \, dz(i)
\]

\[
= \sum_{i=1}^{m} \mathbb{E}_{z(i) \sim Q_i} \left[ \log p(x(i) \mid z(i); \mu, \Lambda, \Psi) + \log p(z(i)) - \log Q_i(z(i)) \right]
\]

\[
= \sum_{i=1}^{m} \mathbb{E}_{z(i) \sim Q_i} \left[ \log \frac{1}{(2\pi)^{n/2} |\Psi|^{1/2}} \exp \left( -\frac{(x(i) - \mu - \Lambda z(i))^T \Psi^{-1} (x(i) - \mu - \Lambda z(i))}{2} \right) + \log p(z(i)) - \log Q_i(z(i)) \right]
\]

\[
= \sum_{i=1}^{m} \mathbb{E}_{z(i) \sim Q_i} \left[ -\frac{1}{2} \log |\Psi| - \frac{n}{2} \log(2\pi) - \frac{1}{2} (x(i) - \mu - \Lambda z(i))^T \Psi^{-1} (x(i) - \mu - \Lambda z(i)) + \log p(z(i)) - \log Q_i(z(i)) \right]
\]
Let

\[
\nabla \Lambda \sum_{i=1}^{m} - \mathbb{E}\left[\frac{1}{2}(x^{(i)} - \mu - \Lambda z^{(i)})^T \psi^{-1}(x^{(i)} - \mu - \Lambda z^{(i)})\right]
\]

\[
= \sum_{i=1}^{m} \nabla \Lambda \mathbb{E}_{z^{(i)} \sim Q_i} \left[ - \text{tr} \left( \frac{1}{2} z^{(i)} T \Lambda^T \psi^{-1} \Lambda z^{(i)} \right) + \text{tr} \left( z^{(i)} T \Lambda^T \psi^{-1} (x^{(i)} - \mu) \right) \right]
\]

\[
= \sum_{i=1}^{m} \nabla \Lambda \mathbb{E}_{z^{(i)} \sim Q_i} \left[ - \text{tr} \left( \frac{1}{2} \Lambda^T \psi^{-1} \Lambda z^{(i)} z^{(i)} T \right) + \text{tr} \left( \Lambda^T \psi^{-1} (x^{(i)} - \mu) z^{(i)} T \right) \right]
\]

\[
= \sum_{i=1}^{m} \mathbb{E}_{z^{(i)} \sim Q_i} \left[ - \psi^{-1} \Lambda z^{(i)} z^{(i)} T + \psi^{-1} (x^{(i)} - \mu) z^{(i)} T \right]
\]

\[
= 0
\]

we have

\[
\Lambda = \left( \sum_{i=1}^{m} (x^{(i)} - \mu) \mathbb{E}_{z^{(i)} \sim Q_i} \left[ z^{(i)} T \right] \right) \left( \sum_{i=1}^{m} \mathbb{E}_{z^{(i)} \sim Q_i} \left[ z^{(i)} z^{(i)} T \right] \right)^{-1}
\]

\[
= \left( \sum_{i=1}^{m} (x^{(i)} - \mu) \mu_{z^{(i)} | x^{(i)}} T \right) \left( \sum_{i=1}^{m} \mu_{z^{(i)} | x^{(i)}} \mu_{z^{(i)} | x^{(i)}} T \right) + \sum_{z^{(i)} | x^{(i)}} \right)^{-1}
\]
EM Algorithm for Factor Analysis (Contd.)

- 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 $\mu$ and $\Psi$

- Results are as follows

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

$$\Psi = \text{diag}(\frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)^T} - x^{(i)} \mu_{z^{(i)}|x^{(i)}} \Lambda^T - \Lambda \mu_{z^{(i)}|x^{(i)} x^{(i)^T} + \Lambda(\mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \sum_{z^{(i)}|x^{(i)}} \Lambda^T)$$
Thanks!

Q & A