# EM Algorithm for Maximum Likelihood Estimation of the Factor Model

Dylan Dijk

# Outline

# Factor Model Recap

$X$ is is an $\mathbb{R}^p$ valued random variable.

Factor Models aim to explain the correlation between variables via a small number of $k < p$ factors.

$$X = \mathbf{\Lambda} F + U$$

# Factor Model Recap

$X$ is is an $\mathbb{R}^p$ valued random variable.

Factor Models aim to explain the correlation between variables via a small number of $k < p$ factors.

$$X = \mathbf{\Lambda} F + U$$

- $\underset{(p \times k)}{\mathbf{\Lambda}}$ is the **loadings matrix** of constants.

# Factor Model Recap

$X$ is is an $\mathbb{R}^p$ valued random variable.

Factor Models aim to explain the correlation between variables via a small number of $k < p$ factors.

$$X = \boldsymbol{\Lambda} F + U$$

- $\underset{(p \times k)}{\boldsymbol{\Lambda}}$ is the **loadings matrix** of constants.
- $F$ is an $\mathbb{R}^k$ valued random variable, called the **factor**.
  - $\mathbb{E}[F] = 0 \qquad \text{Var}(F) = \boldsymbol{I}_k$

# Factor Model Recap

$X$ is is an $\mathbb{R}^p$ valued random variable.

Factor Models aim to explain the correlation between variables via a small number of $k < p$ factors.

$$X = \mathbf{\Lambda} F + U$$

- $\underset{(p \times k)}{\mathbf{\Lambda}}$ is the **loadings matrix** of constants.
- $F$ is an $\mathbb{R}^k$ valued random variable, called the **factor**.
  - $\mathbb{E}[F] = 0 \quad \text{Var}(F) = \boldsymbol{I}_k$
- $U$ is an $\mathbb{R}^p$ valued random variable
  - $\mathbb{E}[U] = 0 \quad \text{Var}(U) = \mathbf{\Psi} = \text{diag}(\psi_{11}, \ldots, \psi_{pp})$

$$\text{Cov}(F, U) = 0$$

# Factor Model Recap

- $\text{Var}(X) = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$
- $\text{Var}(X|F) = \mathbf{\Psi}$
- In lectures we used the **iterated principal factor analysis algorithm** to estimate $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ from the correlation matrix $\mathbf{R}$.

# Additional Assumptions

If we make probabilistic assumptions for $X$ and $F$ we can then use MLE estimates.

- $X|F \sim N_p(\mathbf{\Lambda} F, \mathbf{\Psi})$
- $F \sim N_k(0, \boldsymbol{I}_k)$

# Additional Assumptions

If we make probabilistic assumptions for $X$ and $F$ we can then use MLE estimates.

- $X|F \sim N_p(\mathbf{\Lambda} F, \mathbf{\Psi})$
- $F \sim N_k(0, \mathbf{I}_k)$

$$\implies \begin{pmatrix} F \\ X \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{I}_k & \mathbf{\Lambda}^T \\ \mathbf{\Lambda} & \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \end{pmatrix} \right)$$

$$\implies X \sim N(0, \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi})$$

# Objective

$$X \sim N(0, \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi})$$

Given a dataset $\mathbf{X}$, where rows are i.i.d copies of $X$, we want to estimate the parameters.

$$L(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{X}) = \sum_{i=1}^{n} log(N(x_i; 0, \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}))$$

No closed form solution for finding MLE estimates of $\mathbf{\Psi}$ and $\mathbf{\Lambda}$. Can use the EM algorithm.

# EM Algorithm

- The EM algorithm is a very general technique for finding MLE solutions for probabilistic models with latent variables.
- Latent variables, $Z$, are variables that are not observed.
- EM algorithm is used in cases where direct optimisation of $L(\theta, \mathbf{X}) := log(p(\mathbf{X}; \theta))$ is difficult, but the optimisation of $log(p(\mathbf{X}, \mathbf{Z}; \theta))$ is much easier.

# EM Algorithm - Setup

- Complete-data likelihood $p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$
- Incomplete-data likelihood $p(\mathbf{X}; \boldsymbol{\theta})$

We do not have the complete-data likelihood, so the idea is to maximise its expectation instead.

# EM Algorithm

- **E-Step**

  Compute the latent variable posteriors $p(z|x_i;\theta^{\text{old}})$

- **M-Step**

$$\begin{aligned}\theta^{new} &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \int_z p(z|x_i;\theta^{\text{old}}) \log p(x_i, z; \theta)\; dz \\ &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \mathbb{E}_{Z|X=x_i;\theta^{\text{old}}}[log(p(x_i, z; \theta))]\end{aligned}$$

# EM Algorithm

The EM algorithm will increase the likelihood function $L(\theta, \mathbf{X})$

Let $q(\cdot)$ be a density over the latent variables Z. Then we can decompose the likelihood for a single observed value as:

$$L(\theta, x) = \mathcal{L}(q, \theta) \; + \; \mathrm{KL}(q||p)$$

# EM Algorithm

The EM algorithm will increase the likelihood function $L(\theta, \mathbf{X})$

Let $q(\cdot)$ be a density over the latent variables Z. Then we can decompose the likelihood for a single observed value as:

$$L(\theta, x) = \mathcal{L}(q, \theta) \ + \ \mathrm{KL}(q||p)$$

$$\mathcal{L}(q, \theta) = \int_z q(z) log \left\{ \frac{p(x, z; \theta)}{q(z)} \right\} \ dz$$

$$\mathrm{KL}(q||p) = -\int_z q(z) log \left\{ \frac{p(z|x; \theta)}{q(z)} \right\} \ dz$$

# EM Algorithm

$$L(\theta, x) = \mathcal{L}(q, \theta) \; + \; \mathrm{KL}(q||p)$$

$$\mathrm{KL}(q||p) \geq 0 \implies L(\theta, x) \geq \mathcal{L}(q, \theta)$$

# EM Algorithm

$$L(\theta, x) = \mathcal{L}(q, \theta) \ + \ \mathrm{KL}(q||p)$$

$$\mathrm{KL}(q||p) \geq 0 \implies L(\theta, x) \geq \mathcal{L}(q, \theta)$$

- **E-Step**

Fixing a starting value of the parameters $\theta^{\text{old}}$, $\mathcal{L}(q, \theta)$ is maximised with respect to $q$

$$\underset{q}{\operatorname{argmax}} \, \mathcal{L}(q, \theta^{\text{old}}) = p(z|x; \theta^{\text{old}})$$

# EM Algorithm

$$L(\theta, x) = \mathcal{L}(q, \theta) \; + \; \mathrm{KL}(q||p)$$

$$\mathrm{KL}(q||p) \geq 0 \implies L(\theta, x) \geq \mathcal{L}(q, \theta)$$

- **E-Step**

Fixing a starting value of the parameters $\theta^{\text{old}}$, $\mathcal{L}(q, \theta)$ is maximised with respect to $q$

$$\underset{q}{\operatorname{argmax}} \, \mathcal{L}(q, \theta^{\text{old}}) = p(z|x; \theta^{\text{old}})$$

- **M-Step**

Now keeping $q$ fixed and maximising with respect to $\theta$

$$\underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(q, \theta) = \underset{\theta}{\operatorname{argmax}} \, \mathbb{E}_{Z|X;\theta^{\text{old}}}[log(p(x, z; \theta))]$$

# EM Steps for MLE of the Factor Model

Our model is constructed with a latent variable $F$, therefore we can use the EM algorithm with $F$ in place of $Z$.

- $X|F \sim N_p(\boldsymbol{\Lambda} F, \boldsymbol{\Psi})$
- $F \sim N_k(0, \boldsymbol{I}_k)$

$$\begin{pmatrix} F \\ X \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{\Lambda}^T \\ \boldsymbol{\Lambda} & \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \end{pmatrix} \right)$$

# EM Steps for MLE of the Factor Model

- **E-Step** Compute $p(Z_i|X_i;\boldsymbol{\theta}^{\text{old}})$

$$
\begin{aligned}
\mu_{F_i|X_i} &= \boldsymbol{\Lambda}^T(\boldsymbol{\Psi}+\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)^{-1}(X_i) \\
\Sigma_{F_i|X_i} &= \boldsymbol{I}_k - \boldsymbol{\Lambda}^T(\boldsymbol{\Psi}+\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)^{-1}\boldsymbol{\Lambda}
\end{aligned}
$$

- **M-Step**

$$
\begin{aligned}
\boldsymbol{\Lambda}^{\text{new}} &= \left(\sum_{i=1}^{n}(X_i)\mathbb{E}[F_i]^T\right)\left(\sum_{i=1}^{n}\mathbb{E}[F_iF_i^T]\right)^{-1} \\
\Psi^{\mathrm{new}} &= \frac{1}{n}\text{diag}\left\{\sum_{i=1}^{n}\left(X_iX_i^T - \boldsymbol{\Lambda}^{\mathrm{new}}\mathbb{E}[F_i]X_i\right)\right\}
\end{aligned}
$$

# References

Bartholomew, David J., Martin Knott, and Irini Moustaki (2011). "Latent Variable Models and Factor Analysis: A Unified Approach, 3rd Edition Wiley". In: *Wiley.com*.

Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. URL: https://link.springer.com/book/9780387310732.

# Derivation of EM steps for Factor Model

- **E-Step** Compute $p(Z_i | X_i; \boldsymbol{\theta}^{\text{old}})$

  Immediate from Gaussian identities

- **M-Step**

$$\theta^{new} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbb{E}_{Z_i | X_i; \boldsymbol{\theta}^{\text{old}}} [log(p(X_i, Z_i; \theta)]$$

In our case we have that

$$p(X_i, F_i; \theta) = p(X_i | F_i; \theta) p(F_i)$$

$p(F_i)$ does not depend on our parameters of interest.

# Derivation of EM steps for Factor Model

$$Q := \sum_{i=1}^{n} \mathbb{E}_{F_i|X_i;\boldsymbol{\theta}^{\text{old}}}[log(p(X_i|F_i;\theta)]$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\log\left((2\pi)^{p/2}|\Psi|^{-1/2}\exp\left\{-\frac{1}{2}[X_i - \boldsymbol{\Lambda}F_i]^T\Psi^{-1}[X_i - \boldsymbol{\Lambda}F_i]\right\}\right)\right]$$

$$= c - \frac{n}{2}\log|\Psi|$$

$$-\sum_{i=1}^{n}\left(\frac{1}{2}X_i^T\boldsymbol{\Psi}^{-1}X_i - X_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\mathbb{E}[F_i] + \frac{1}{2}\left[\mathbb{E}[F_i^T\boldsymbol{\Lambda}^T\Psi^{-1}\boldsymbol{\Lambda}F_i]\right]\right)$$

# M-Step

Now need to maximise with respect to $\theta = (\boldsymbol{\Lambda}, \boldsymbol{\Psi})$

$$\frac{\partial Q}{\partial \boldsymbol{\Lambda}} = \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial \boldsymbol{\Lambda}} X_i^T \Psi^{-1} \boldsymbol{\Lambda} \mathbb{E}[F_i]}_{(1)} - \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\Big[ \underbrace{\frac{\partial}{\partial \boldsymbol{\Lambda}} F_i^T \boldsymbol{\Lambda}^T \Psi^{-1} \boldsymbol{\Lambda} F_i}_{(2)} \Big]$$

$$\begin{aligned} (1) &= \mathbb{E}[F_i](X_i^T \Psi^{-1}) \\ (2) &= (2\Psi^{-1} \boldsymbol{\Lambda} F_i F_i^T)^T \end{aligned}$$

Setting to zero we get:

$$\boldsymbol{\Lambda}^{\text{new}} = \left( \sum_{i=1}^{n} (X_i) \mathbb{E}[F_i]^T \right) \left( \sum_{i=1}^{n} \mathbb{E}[F_i F_i^T] \right)^{-1}$$

# M-Step

$$\frac{\partial Q}{\partial \boldsymbol{\Psi}^{-1}} = \frac{n}{2}\Psi^{\text{new}} - \sum_{i=1}^{n}\left(\frac{1}{2}X_iX_i^T - \boldsymbol{\Lambda}^{\text{new}}\mathrm{E}[F_i]X_i^T + \frac{1}{2}\boldsymbol{\Lambda}^{\text{new}}\mathrm{E}[F_iF_i^T](\boldsymbol{\Lambda}^{\text{new}})^T\right)$$

Setting to zero and plugging in $\boldsymbol{\Lambda}^{\text{new}}$:

$$\frac{n}{2}\Psi^{\text{new}} = \sum_{i=1}^{n}\left(\frac{1}{2}X_iX_i^T - \frac{1}{2}\boldsymbol{\Lambda}^{\text{new}}\mathbb{E}[F_i]X_i\right)$$

Restricting to a diagonal matrix $\boldsymbol{\Psi}$:

$$\Psi^{\text{new}} = \frac{1}{n}\text{diag}\left\{\sum_{i=1}^{n}\left(X_iX_i^T - \boldsymbol{\Lambda}^{\text{new}}\mathbb{E}[F_i]X_i\right)\right\}$$