Linear Models: FA, ICA, NFA

Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

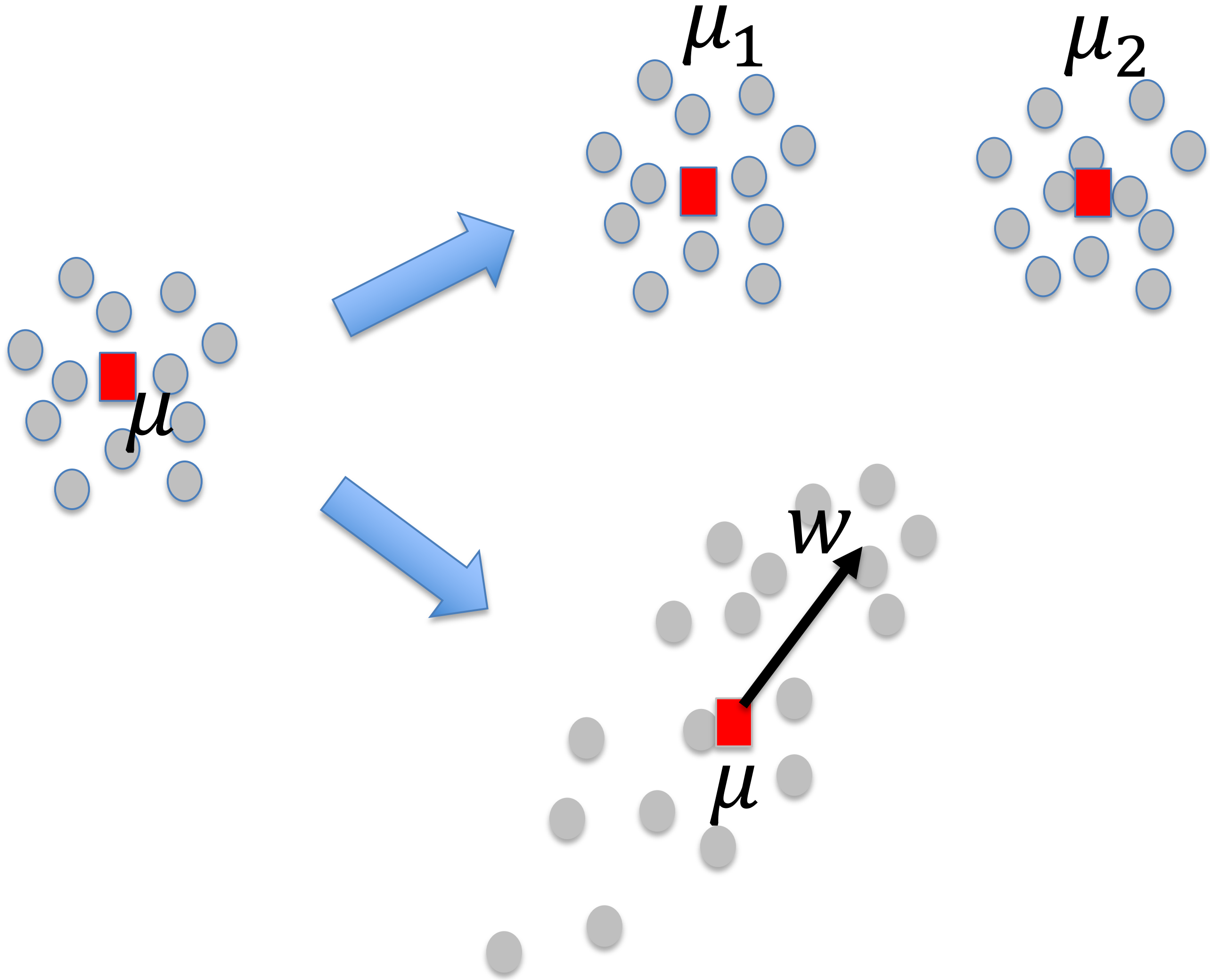
2021-04-20

Outline

- **Recall**
  - **Principal Component Analysis (PCA)**
  - **Hebbian learning, Oja's, LMSER and PCA**

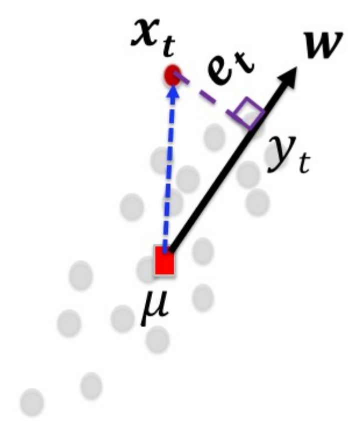
- Probabilistic PCA, Factor Analysis (FA)

Model from "one point" to "one line"



PCA by minimizing MSE

$$J(w) = \frac{1}{N}\sum_{t=1}^{N} ||\boldsymbol{x}_t - (\boldsymbol{x}_t^T \boldsymbol{w})\boldsymbol{w}||^2$$



$$\boldsymbol{x}_t^T\boldsymbol{x}_t - (\boldsymbol{x}_t^T\boldsymbol{w})\boldsymbol{w}^T\boldsymbol{x}_t - \boldsymbol{x}_t^T(\boldsymbol{x}_t^T\boldsymbol{w})\boldsymbol{w} + (\boldsymbol{x}_t^T\boldsymbol{w})\boldsymbol{w}^T(\boldsymbol{x}_t^T\boldsymbol{w})\boldsymbol{w}$$
$$= \boldsymbol{x}_t^T\boldsymbol{x}_t - \boldsymbol{w}^T(\boldsymbol{x}_t\boldsymbol{x}_t^T)\boldsymbol{w}$$

$$||w|| = 1$$

Introduce a Lagrange multiplier λ

$$L(\{\boldsymbol{x}_t\}, \boldsymbol{w}) = J(\{\boldsymbol{x}_t\}, \boldsymbol{w}) - \lambda \cdot (\boldsymbol{w}^T\boldsymbol{w} - 1)$$

$$\Sigma_x = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{x}_t\boldsymbol{x}_t^T$$

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} - \lambda \cdot \frac{\partial(\boldsymbol{w}^T\boldsymbol{w} - 1)}{\partial \boldsymbol{w}} = -2(\Sigma_x\boldsymbol{w}) - \lambda \cdot 2\boldsymbol{w} = \boldsymbol{0}$$

$$\boxed{\Sigma_x\boldsymbol{w} = (-\lambda)\cdot\boldsymbol{w}}$$

Eigenvalues and Eigenvectors

Algorithms for PCA

$$\Sigma_x = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^T$$

- Eigen-decomposition $\boxed{\Sigma_x \boldsymbol{w} = (-\lambda) \cdot \boldsymbol{w}}$
- SVD $X = UDV^T \qquad XX^T = UDV^T \cdot VDU^T = UD^2U^T$
- Hebbian learning rule $\tau^W \dfrac{dW}{dt} = \bar{z}\bar{x}^t$
- Oja learning rule $\tau^W \dfrac{dW}{dt} = \bar{z}\bar{x}^t - \bar{y}\bar{u}^t$
- Lmser rule $\tau^W \dfrac{dW}{dt} = \bar{z}\bar{x}^t - \bar{y}\bar{u}^t + \bar{z}\bar{x}^t - \bar{y}^t\bar{x}^t$

$\vec{z} = \vec{y} \qquad \vec{y} = W\vec{x}, \vec{u} = W^t\vec{y}, \vec{y}^r = W\vec{u}$

Outline

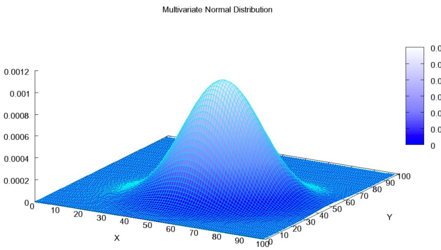
- Recall
  - Principal Component Analysis (PCA)
  - Hebbian learning, Oja's, LMSER and PCA

- **Probabilistic PCA, Factor Analysis (FA)**

Gaussian distributions

Density function

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$



Bivariate Gaussian:

In the 2-dimensional nonsingular case ($k = rank(\sum) = 2$), the probability density function of a vector $[XY]'$ is:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

where ρ is the correlation between X and Y and where $\sigma_X > 0$ and $\sigma_Y > 0$. In this case,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Conditional Gaussian

If N-dimensional $\mathbf{x}$ is partitioned as follows

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

and accordingly $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as follows

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

then the distribution of $\mathbf{x}_1$ conditional on $\mathbf{x}_2 = \mathbf{a}$ is multivariate normal $(\mathbf{x}_1 \mid \mathbf{x}_2 = \mathbf{a}) \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ where

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_2\right)$$

and covariance matrix

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Affine transformation

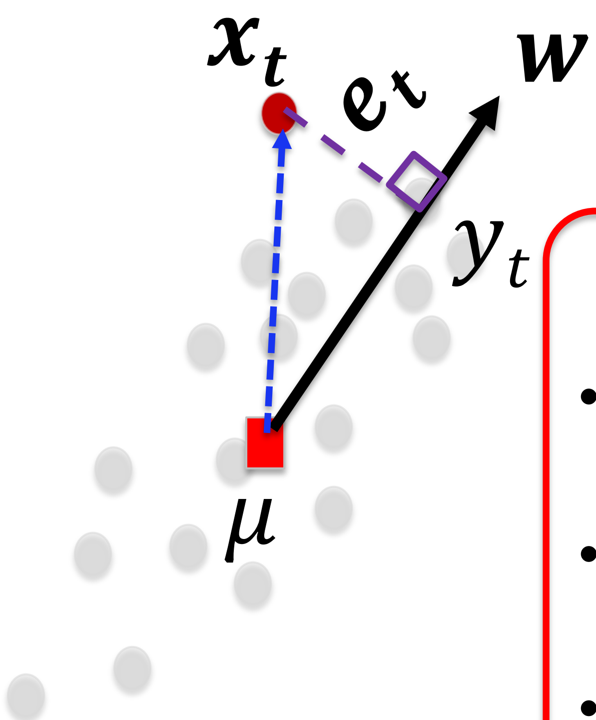
If $\mathbf{Y} = \mathbf{c} + \mathbf{BX}$ is an affine transformation of $\mathbf{X} \sim \mathcal{N}(\boldsymbol\mu, \boldsymbol\Sigma)$, where $\mathbf{c}$ is an $M \times 1$ vector of constants and $\mathbf{B}$ is a constant $M \times N$ matrix, then $\mathbf{Y}$ has a multivariate normal distribution with expected value $\mathbf{c} + \mathbf{B}\boldsymbol\mu$ and variance $\mathbf{B\Sigma B}^{\mathrm{T}}$ i.e., $\mathbf{Y} \sim \mathcal{N}\left(\mathbf{c} + \mathbf{B}\boldsymbol\mu, \mathbf{B\Sigma B}^{\mathrm{T}}\right)$.

If $\boldsymbol\Sigma = \mathbf{U\Lambda U}^{\mathrm{T}} = \mathbf{U\Lambda}^{1/2}(\mathbf{U\Lambda}^{1/2})^{\mathrm{T}}$ is an eigendecomposition where the columns of $\mathbf{U}$ are unit eigenvectors and $\boldsymbol\Lambda$ is a diagonal matrix of the eigenvalues, then we have

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol\mu, \boldsymbol\Sigma) \iff \mathbf{X} \sim \boldsymbol\mu + \mathbf{U\Lambda}^{1/2}\mathcal{N}(0, \mathbf{I}) \iff \mathbf{X} \sim \boldsymbol\mu + \mathbf{U}\mathcal{N}(0, \boldsymbol\Lambda).$$

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

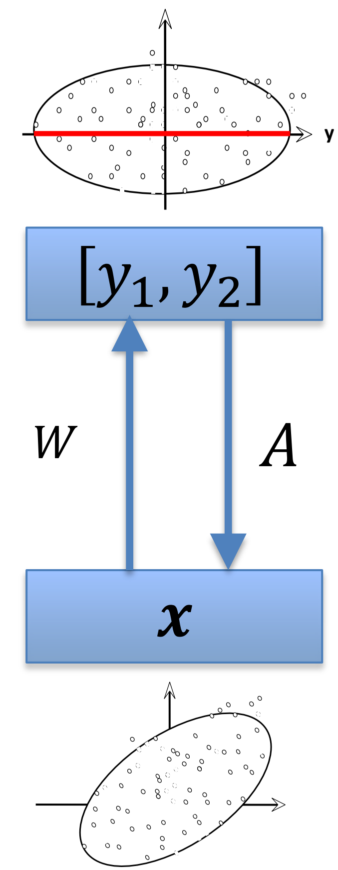
Generative model perspective



Continuous latent variable $\boldsymbol{y}$

For the t-th data point:

- Randomly sample a y_t:
$$\boldsymbol{y}_t \sim G(\boldsymbol{y}|\boldsymbol{0}, \Sigma_y);$$
- Randomly generate a noise e_t
$$\boldsymbol{e}_t \sim G(\boldsymbol{e}|0, \sigma^2 I)$$
- Generate $\boldsymbol{x}_t$ by:
$$\boldsymbol{x}_t = A\boldsymbol{y}_t + \boldsymbol{\mu} + \boldsymbol{e}_t$$



$$||\boldsymbol{w}|| = 1$$
$$y_t = \boldsymbol{x}_t^T \boldsymbol{w}$$
$$\boldsymbol{e_t} = ||\boldsymbol{x_t} - y_t \boldsymbol{w}||^2$$

$\boldsymbol{y}_t$ and $\boldsymbol{e}_t$ are independent

An illustration of the generative view

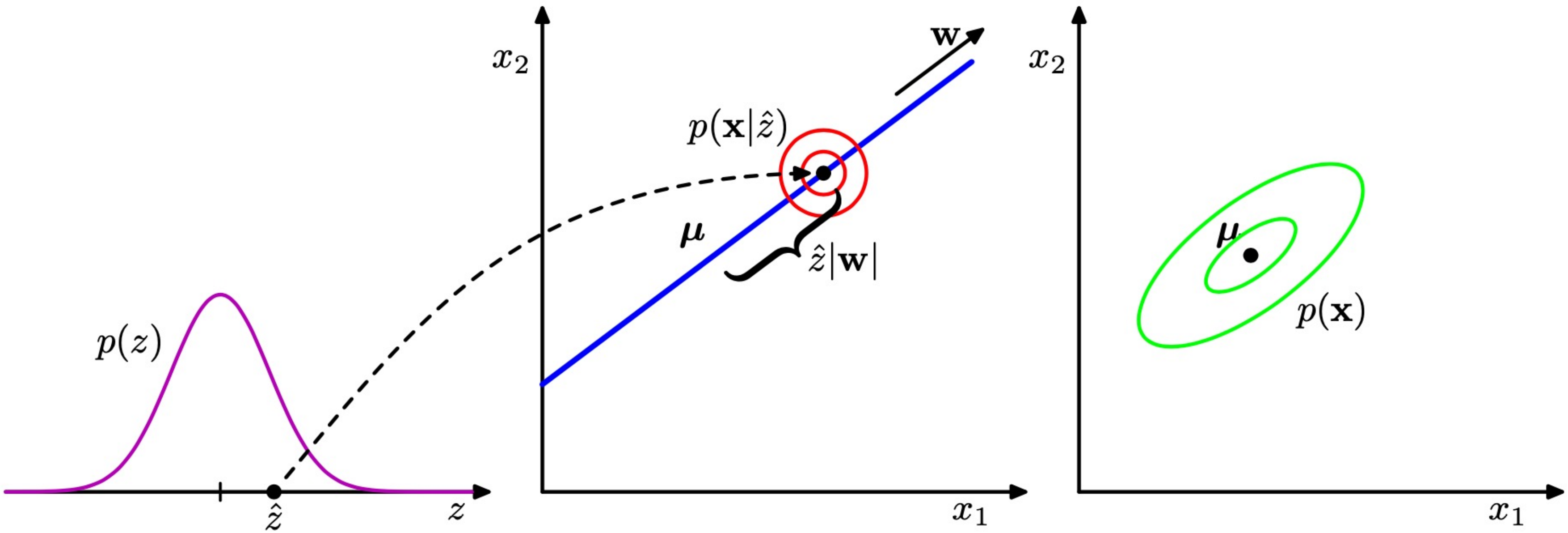
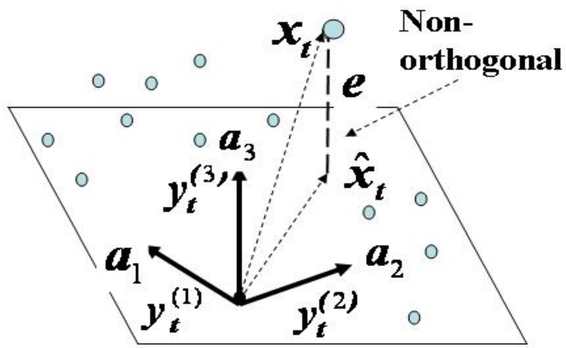


Figure 12.9 An illustration of the generative view of the probabilistic PCA model for a two-dimensional data space and a one-dimensional latent space. An observed data point $\mathbf{x}$ is generated by first drawing a value $\widehat{z}$ for the latent variable from its prior distribution $p(z)$ and then drawing a value for $\mathbf{x}$ from an isotropic Gaussian distribution (illustrated by the red circles) having mean $\mathbf{w}\widehat{z} + \boldsymbol{\mu}$ and covariance $\sigma^2\mathbf{I}$. The green ellipses show the density contours for the marginal distribution $p(\mathbf{x})$.

PRML book, P572

Factor Analysis (FA) Model

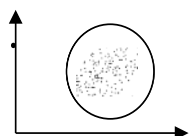


$A^T A = I$ has been removed because it impedes $\sum_t \|e_t\|^2$ to reach its minimum

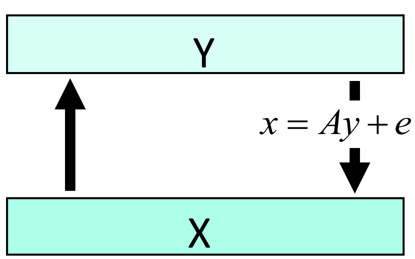
Assumption

$c)$ $Eey^T = 0$ (i.e., not correlated) it is not longer a consequence;

$f)$ $Eyy^T = \Lambda$ is diagonal



$$q(y) = \prod_{j=1}^{k} G(y_j|0,1) = G(y|0,I)$$



Indeterminacy

$e)$ a rotation matrix since $A' = \Phi A$ spans the same subspace;

$f)$ a diagonal D with $A' = AD$, $y' = D^{-1} y$.

Two Choices

$$q(y) = \begin{cases} G(y|0,I), & \text{choice (a)} \\ G(y|0,\Lambda), & \text{choice (b)} \end{cases}$$

g) a unknown allocation between the two additive terms $Exx^T = A^T \Lambda A + Eee^T$.

- 样本方差之分割不变性（样本方差在子空间内外可任意分割）
- 超阈维数不变性（高于某维的子空间以零误差描述有限样本集）

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

   and return to step 2.

EM algorithm for FA

E-Step: $$p^{old}(\boldsymbol{y}|\boldsymbol{x}) = \frac{G(\boldsymbol{y}|0, I)G(\boldsymbol{x}|A\boldsymbol{y} + \boldsymbol{\mu}, \sigma^2 I)}{G(\boldsymbol{x}|\boldsymbol{\mu}, AA^T + \sigma^2 I)}$$

$$E[\boldsymbol{y}|\boldsymbol{x}] = W\boldsymbol{x} \qquad W = A^T(AA^T + \sigma^2 I)^{-1}$$

$$E[\boldsymbol{y}\boldsymbol{y}^T|\boldsymbol{x}] = I - WA + W\boldsymbol{x}\boldsymbol{x}^T W^T$$

M-Step: max $Q\left(p^{old}(\boldsymbol{y}|\boldsymbol{x}), \Theta\right)$

$$Q = \int p^{old}(\boldsymbol{y}|\boldsymbol{x}) \cdot \ln[G(\boldsymbol{y}|0, I)G(\boldsymbol{x}|A\boldsymbol{y} + \boldsymbol{\mu}, \sigma^2 I)]\, d\boldsymbol{y}$$

$$A^{new} = \left(\sum_{t=1}^{N} \boldsymbol{x}_t (E[\boldsymbol{y}|\boldsymbol{x}_t])^T\right)\left(\sum_{t=1}^{N} E[\boldsymbol{y}\boldsymbol{y}^T|\boldsymbol{x}_t]\right)^{-1}$$

$$\sigma^{2\,new} = \frac{1}{Nd} Tr\left\{\sum_{t=1}^{N} \{\boldsymbol{x}_t\boldsymbol{x}_t^T - A^{new}E[\boldsymbol{y}|\boldsymbol{x}_t]\boldsymbol{x}_t^T\}\right\}$$

Maximum likelihood FA implements PCA

$$p(\boldsymbol{y}) = G(\boldsymbol{y}|\mathbf{0}, I), \qquad p(\boldsymbol{x}|\boldsymbol{y}) = G(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{\mu}, \Sigma_e),$$

$$p(\boldsymbol{x}|\Theta) = \int p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{y} = G\big(\boldsymbol{x}\big|\boldsymbol{\mu}, \mathrm{AA}^{\mathrm{T}} + \Sigma_e\big),$$

$$\max_{\Theta} \log\left\{\prod_{t=1}^{N=1} p(\boldsymbol{x}_t|\Theta)\right\}$$

Maximum Likelihood

$$\boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}_n$$

assume $\boldsymbol{\mu} = \mathbf{0}$.

PCA
$$\begin{cases} \hat{\mathbf{A}}_{n\times m}^{ML} = \mathbf{U}_{n\times m}(\mathbf{D}_m - \hat{\sigma}_e^2)^{\frac{1}{2}}\mathbf{R}^T, \quad \mathbf{D}_m = diag[s_1, \ldots, s_m] \\ \hat{\sigma}_e^{2,ML} = \frac{1}{n-m}\sum_{i=m+1}^{n} s_i, \end{cases}$$

U is eigenvectors of sample cov.

Thank you!