Latent Variable Models:

Factor Analyzers and Mixture of Factor Analyzers

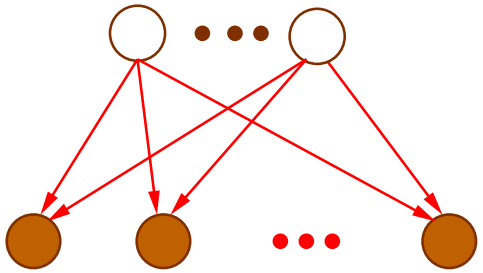
Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Outline

- Why latent variable models?
- Principal component analysis (PCA)
- Maximum likelihood factor analysis
- Probabilistic PCA
- Mixture of factor analyzers

Why Latent Variable Models?: An Example



Gaussian $p(x_1, \ldots, x_D)$ requires D independent parameters for mean vector and $\frac{D(D+1)}{2}$ independent parameters for covariance matrix, $\frac{D(D+3)}{2}$ parameters in total.

The number of independent parameters grows with D^2.

Marginal independence assumes:

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i)$$

which requires just $2D$ free parameters.

Conditional independence assumes:

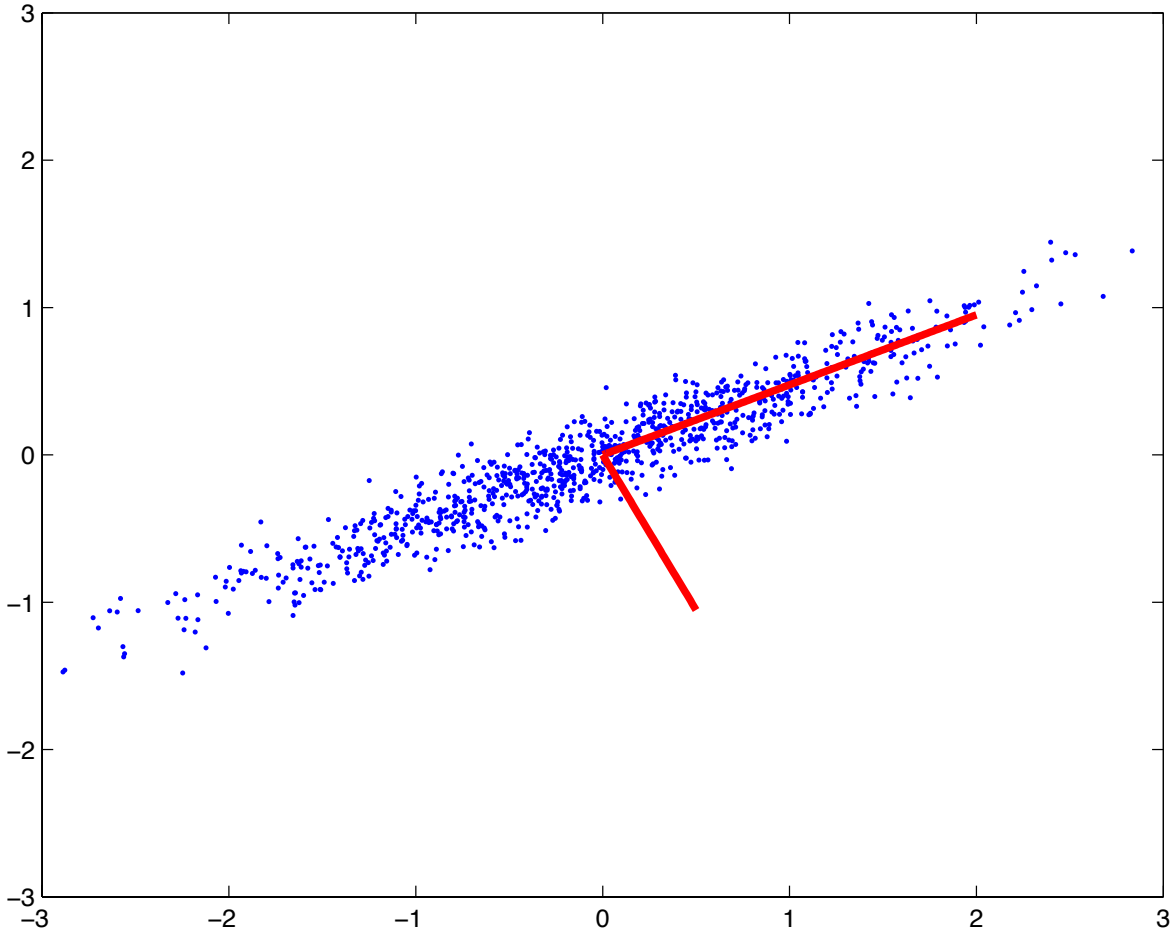
$$p(x_1, \ldots, x_D | s_1, \ldots, s_K) = \prod_{i=1}^{D} p(x_i | s_1, \ldots, s_K).$$

In the case of linear models, the number of independent parameters grows with D (actually, need $(DK + 2D)$).

Principal Component Analysis (PCA)

- Principal component analysis (PCA) is a well-established technique for dimension reduction. Its applications include data compression, image processing, data visualization, exploratory data analysis, pattern recognition, and time series prediction.
- The most common derivation of PCA is in terms of an orthogonal projection ($\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}$) which maximizes the variance in the projected space. Given a set of m-dimensional observation vector, $\{\boldsymbol{x}_t\}$, the PCA aims at finding a orthogonal linear projection $\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}$ such that the variance of $\boldsymbol{y} \in \mathbb{R}^d$ $(d < D)$ is maximized. The ith element of $\boldsymbol{y}$ is called *ith principal component*.
- Alternatively, PCA provides an orthogonal linear projection which minimize the squared reconstruction error $\sum_{t=1}^{N} \|\boldsymbol{x}_t - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{x}_t\|^2$. Thus PCA is the optimal linear encoding in MS sense.
- It was shown that $\boldsymbol{w}_i$ (ith column vector of $\boldsymbol{W}$) corresponds to the normalized eigenvector associated with the ith largest eigenvalue of the covariance matrix $\boldsymbol{C}_x = \mathbb{E}\{(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\}$.

PCA: An Example



Vanilla PCA

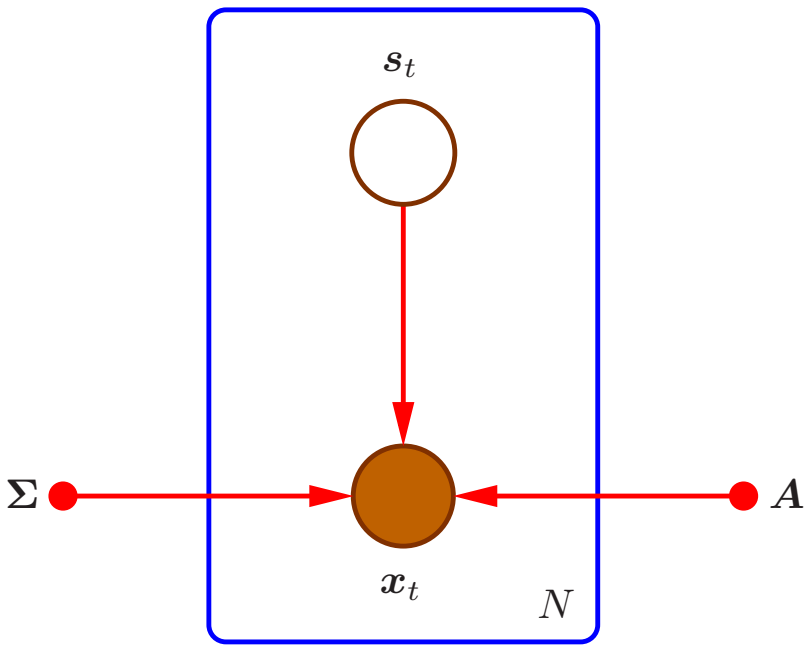
- Data centering: $\boldsymbol{X} \leftarrow \boldsymbol{X}\left(\boldsymbol{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)$, where $\mathbf{1}_N \in \mathbb{R}^N$ is the vector of all one's.
- Find rank-d approximation of the sample covariance matrix $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{D\times D}$:

$$\boldsymbol{C} \approx \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top,$$

  where $\boldsymbol{U} \in \mathbb{R}^{D\times d}$ contains d leading eigenvectors in its columns and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is diagonal matrix in which diagonal entries are eigenvalues in descending order.
- Principal components are computed as

$$\boldsymbol{Y} = \boldsymbol{U}^\top\boldsymbol{X}.$$

Factor Analysis: Graphical Model



- Gaussian latent variables:

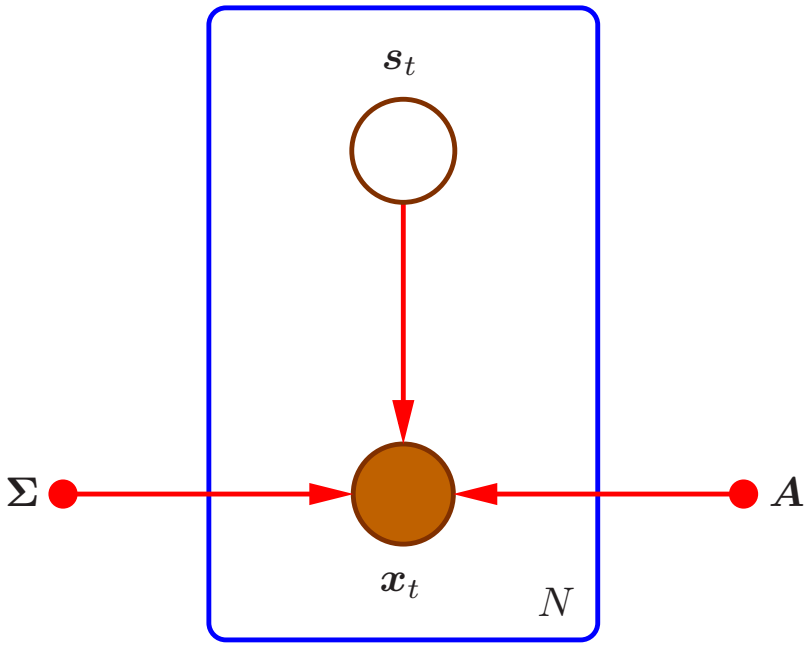
$$p(\boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{s}_t|0, \boldsymbol{I}).$$

- The distribution over observed variables conditioned on latent variables is

$$p(\boldsymbol{x}_t|\boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{A}\boldsymbol{s}_t, \boldsymbol{\Sigma}),$$

  where $\boldsymbol{\Sigma}$ is a diagonal matrix.
- Finds a lower-dimensional projection of high-dimensional data that captures the covariance structure of the data.

Learning Factor Analysis



- Compute maximum likelihood estimates of parameters

$$\left\{ \boldsymbol{A}, \boldsymbol{\Sigma} \right\}.$$

- Compute the posterior distribution over latent variables

$$p(\boldsymbol{s}_t | \boldsymbol{x}_t)$$

which yields lower-dimensional projection.

- We will do this again using expectation maximization.

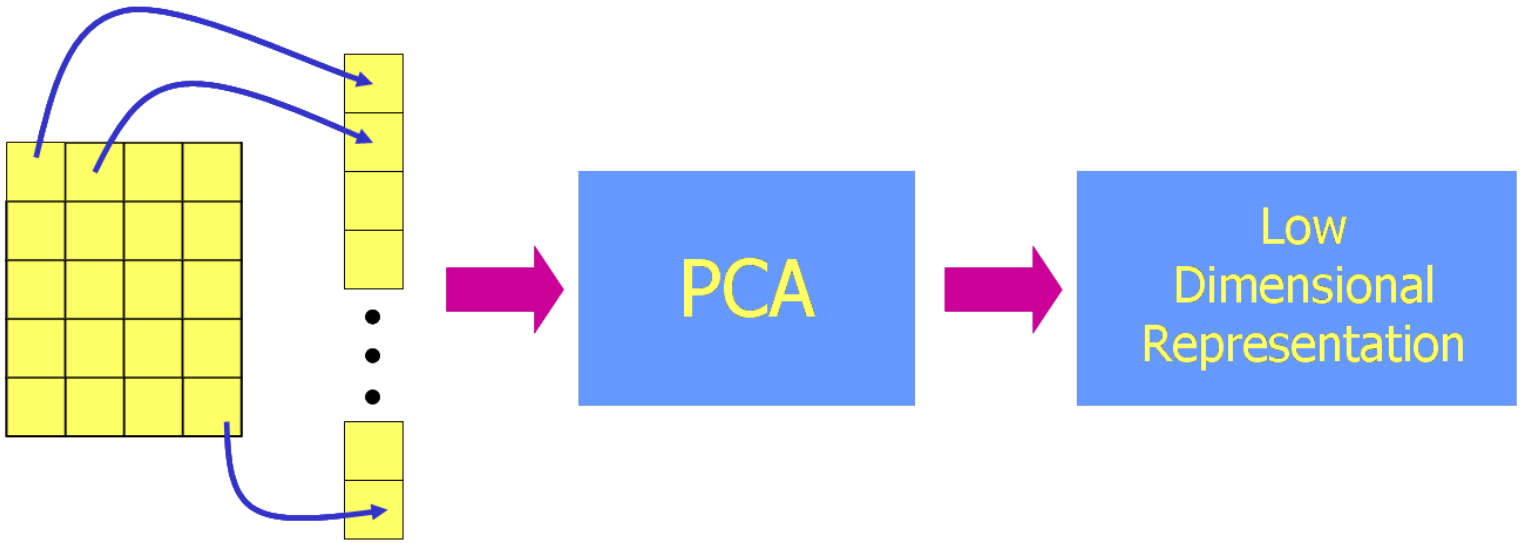
Factor Analysis

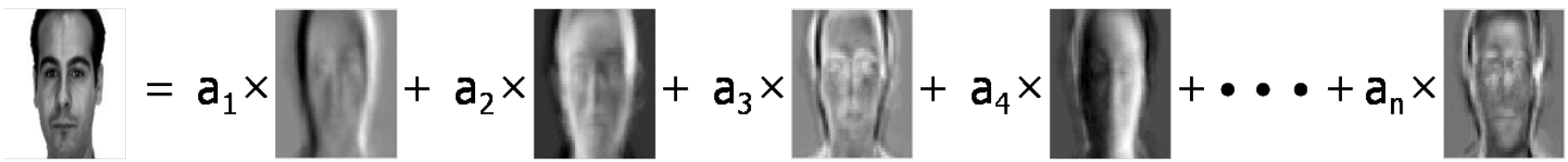
- In maximum likelihood factor analysis (FL), an D-dimensional real-valued data $\boldsymbol{x}$ is modeled using an d-dimensional vector of real-valued factors $\boldsymbol{s}$. In general, d is much smaller than D.
- The linear generative model is given by

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon},$$

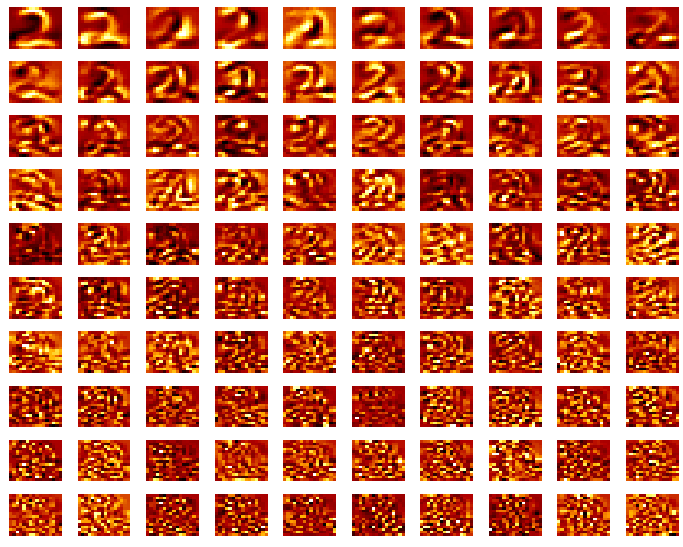
  where $\boldsymbol{A}$ is known as the factor loading matrix. The factors $\boldsymbol{s}$ are assumed to be $\mathcal{N}(0, \boldsymbol{I})$ distributed. The D-dimensional vector $\boldsymbol{\epsilon}$ is $\mathcal{N}(0, \boldsymbol{\Sigma})$ distributed, where $\boldsymbol{\Sigma}$ is a diagonal matrix.
- The observed variables $\boldsymbol{x}$ are independent given the factors. According to this model, $\boldsymbol{x}$ is normally distributed with zero mean and covariance $\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Sigma}$.
- The goal of factor analysis is to find the $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ that best model the covariance structure of $\boldsymbol{x}$.
- There is an indeterminacy in factor analysis, since a factor loading matrix $\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is any orthogonal matrix, produces the same covariance of $\boldsymbol{x}$, using hidden factors given by $\boldsymbol{s}' = \boldsymbol{Q}^\top \boldsymbol{s}$. (Grassmann manifold)

Eigenface





Eigendigits



Remarks on Factor Analysis

- The d factors play the same role as the principal components in PCA. They are informative projections of the data.
- The columns of $\boldsymbol{A}$ span the space of the first d principal components.
- To compute the corresponding eigenvector and eigenvalues explicitly, the data can be projected into this d-dimensional subspace and an ordered orthogonal basis for the covariance in the subspace can be considered.
- For the case of isotropic Gaussian noise, the ML FA boils down to probabilistic PCA.
- In the case of zero noise limit, it is simplified to EM-PCA.

EM Algorithm for Maximum Likelihood Factor Analysis

- E-step: We compute the posterior distribution over latent variables, $p(\boldsymbol{s}|\boldsymbol{x})$ and calculate the expected complete-data log-likelihood $\mathbb{E}\mathcal{L}_c$ where the expectation is taken with respect to the posterior $p(\boldsymbol{s}|\boldsymbol{x})$.

$$\begin{aligned}\mathbb{E}\mathcal{L}_c &= \mathbb{E}\log p(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{A}, \boldsymbol{\Sigma}) \\ &= \mathbb{E}\sum_{t=1}^{N}\log p(\boldsymbol{x}_t, \boldsymbol{s}_t|\boldsymbol{A}, \boldsymbol{\Sigma}).\end{aligned}$$

- M-step: Re-estimate parameters $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ which maximize the expected complete-data log-likelihood:

$$\begin{aligned}\boldsymbol{A} &\leftarrow \arg\max_{\boldsymbol{A}} \mathbb{E}\mathcal{L}_c, \\ \boldsymbol{\Sigma} &\leftarrow \arg\max_{\boldsymbol{\Sigma}} \mathbb{E}\mathcal{L}_c.\end{aligned}$$

Gaussian Identities

A D-dimensional Gaussian density for $\boldsymbol{x}$ is

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} \mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

Define the augmented vector $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{s}^\top]^\top$ which is jointly normal, i.e.,

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{s} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{B} \end{bmatrix}\right).$$

The marginal densities are $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{A})$, $\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{B})$.
The conditional distributions are:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{s}) &= \mathcal{N}\left(\boldsymbol{a} + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{s}-\boldsymbol{b}), \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top\right) \\ p(\boldsymbol{s} \mid \boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{b} + \boldsymbol{C}^\top\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{a}), \boldsymbol{B} - \boldsymbol{C}^\top\boldsymbol{A}^{-1}\boldsymbol{C}\right). \end{aligned}$$

E-Step for ML-FA

It follows from Gaussian identities that the posterior distribution over latent variables is calculated as

$$p(\boldsymbol{s}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{s}|\boldsymbol{\Phi}\boldsymbol{x}, \boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{A}),$$

where $\boldsymbol{\Phi} = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Sigma})^{-1}$.

Since $p(\boldsymbol{s}|\boldsymbol{x})$ is Gaussian, we only need the following sufficient statistics to evaluate the expected complete-data log-likelihood:

$$\begin{aligned}
\mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\} &= \boldsymbol{\Phi}\boldsymbol{x}_t, \\
\mathbb{E}\{\boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t\} &= \mathrm{var}(\boldsymbol{s}_t|\boldsymbol{x}_t) + \mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\}\left[\mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\}\right]^\top \\
&= \boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{A} + \boldsymbol{\Phi}\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Phi}^\top.
\end{aligned}$$

The complete-data log-likelihood is given by

$$\mathcal{L}_c = \log p(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{A}, \boldsymbol{\Sigma}) = \sum_{t=1}^{N} \log p(\boldsymbol{x}_t|\boldsymbol{s}_t) + \sum_{t=1}^{N} \log p(\boldsymbol{s}_t).$$

The expected complete-data log-likelihood is given by

$$\begin{aligned}
\mathbb{E}\mathcal{L}_c &= \sum_{t=1}^{N} \mathbb{E} \log p(\boldsymbol{x}_t|\boldsymbol{s}_t) + \sum_{t=1}^{N} \mathbb{E} \log p(\boldsymbol{s}_t) \\
&\propto -\frac{N}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{t=1}^{N} \Big\{ \boldsymbol{x}_t^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t - 2 \left[\mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\}\right]^\top \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t \\
&\quad + \mathrm{tr}\left(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{A}\, \mathbb{E}\{\boldsymbol{s}_t \boldsymbol{s}_t^\top|\boldsymbol{x}_t\}\right) \Big\}.
\end{aligned}$$

M-Step for ML-FA

M-step involves re-estimating parameters $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ which maximize $\mathbb{E}\mathcal{L}_c$:

$$\begin{aligned} \boldsymbol{A} &\leftarrow \underset{\boldsymbol{A}}{\arg\max}\ \mathbb{E}\mathcal{L}_c, \\ \boldsymbol{\Sigma} &\leftarrow \underset{\boldsymbol{\Sigma}}{\arg\max}\ \mathbb{E}\mathcal{L}_c. \end{aligned}$$

Solve $\frac{\partial \mathbb{E}\mathcal{L}_c}{\partial \boldsymbol{A}} = 0$ and $\frac{\partial \mathbb{E}\mathcal{L}_c}{\partial \boldsymbol{\Sigma}^{-1}} = 0$ for $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$, respectively, to obtain

$$\begin{aligned} \boldsymbol{A}^{\text{new}} &= \left(\sum_{t=1}^{N} \boldsymbol{x}_t \left[\mathbb{E}\{\boldsymbol{s}_t | \boldsymbol{x}_t\} \right]^\top \right) \left(\sum_{t=1}^{N} \mathbb{E}\{\boldsymbol{s}_t \boldsymbol{s}_t^\top | \boldsymbol{x}_t\} \right)^{-1}, \\ \boldsymbol{\Sigma}^{\text{new}} &= \frac{1}{N} \text{diag} \left\{ \sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^\top - \boldsymbol{A}^{\text{new}} \sum_{t=1}^{N} \mathbb{E}\{\boldsymbol{s}_t | \boldsymbol{x}_t\} \boldsymbol{x}_t^\top \right\}. \end{aligned}$$

Algorithm Outline: EM for ML-FA

- E-step: Compute sufficient statistics:

$$\begin{aligned}\mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\} &= \boldsymbol{\Phi}\boldsymbol{x}_t, \\ \mathbb{E}\{\boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t\} &= \boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{A} + \boldsymbol{\Phi}\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{\Phi}^\top,\end{aligned}$$

  where $\boldsymbol{\Phi} = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Sigma})^{-1}$.

- M-step: Re-estimate parameters:

$$\begin{aligned}\boldsymbol{A}^{\text{new}} &= \left(\sum_{t=1}^{N}\boldsymbol{x}_t\left[\mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\}\right]^\top\right)\left(\sum_{t=1}^{N}\mathbb{E}\{\boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t\}\right)^{-1}, \\ \boldsymbol{\Sigma}^{\text{new}} &= \frac{1}{N}\text{diag}\left\{\sum_{t=1}^{N}\boldsymbol{x}_t\boldsymbol{x}_t^\top - \boldsymbol{A}^{\text{new}}\sum_{t=1}^{N}\mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\}\boldsymbol{x}_t^\top\right\}.\end{aligned}$$

Probabilistic PCA (PPCA)

Tipping and Bishop showed that the subspace defined by the columns of the factor loading matrix $\boldsymbol{A}$ corresponds to principal subspace when $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ (isotropic Gaussian).

- E-step: Compute sufficient statistics:

$$\begin{aligned} \mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\} &= \boldsymbol{M}^{-1}\boldsymbol{A}^\top \boldsymbol{x}_t, \\ \mathbb{E}\{\boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t\} &= \sigma^2\boldsymbol{M}^{-1} + \mathbb{E}\{\boldsymbol{s}_t|\boldsymbol{x}_t\}\mathbb{E}\{\boldsymbol{s}_t^\top|\boldsymbol{x}_t^\top\}, \end{aligned}$$

  where $\boldsymbol{M} = \boldsymbol{A}^\top\boldsymbol{A} + \sigma^2\boldsymbol{I}$.

- M-step: Re-estimate parameters $\boldsymbol{A}$ and σ^2:

$$\begin{aligned} \boldsymbol{A} &\leftarrow \left[\sum_{t=1}^{N} \boldsymbol{x}_t \mathbb{E}\{\boldsymbol{s}_t^\top|\boldsymbol{x}_t^\top\}\right]\left[\sum_{t=1}^{N} \mathbb{E}\{\boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t\}\right]^{-1}, \\ \sigma^2 &\leftarrow \frac{1}{ND}\sum_{t=1}^{N}\Big\{\|\boldsymbol{x}_t\|^2 - 2\mathbb{E}\{\boldsymbol{s}_t^\top|\boldsymbol{x}_t^\top\}\boldsymbol{A}^\top\boldsymbol{x}_t \\ &\qquad + \operatorname{tr}\left(\boldsymbol{E}\{\boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t\}\right)\boldsymbol{A}^\top\boldsymbol{A}\Big\}. \end{aligned}$$

EM-PCA: Limiting Case

- In general, the subspace defined by the columns of the factor loading matrix $\boldsymbol{A}$ does not correspond to the principal subspace of the data. However, for the case of isotropic Gaussian noise, i.e., $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, FA behaves like PSA.
- In the case of zero noise limit ($\sigma^2 \to 0$), the inference reduces to simple least square projection:

$$\begin{aligned} \boldsymbol{\Phi} &= \lim_{\sigma^2 \to 0} \boldsymbol{A}^\top \left(\boldsymbol{A}\boldsymbol{A}^\top + \sigma^2 \boldsymbol{I} \right)^{-1} = \left(\boldsymbol{A}^T \boldsymbol{A} \right)^{-1} \boldsymbol{A}^T, \\ p(\boldsymbol{s}|\boldsymbol{x}) &= \mathcal{N}\left(\left(\boldsymbol{A}^\top \boldsymbol{A} \right)^{-1} \boldsymbol{A}^\top \boldsymbol{x}, 0 \right) = \delta\left(\boldsymbol{s} - \left(\boldsymbol{A}^\top \boldsymbol{A} \right)^{-1} \boldsymbol{A}^\top \boldsymbol{x} \right). \end{aligned}$$

For $\boldsymbol{A} \in \mathbb{R}^{D \times K}$ with rank K, left multiplication by $\boldsymbol{A}^\top \left(\boldsymbol{A}\boldsymbol{A}^\top \right)^{-1}$ is exactly equivalent to left multiplication by $\left(\boldsymbol{A}^\top \boldsymbol{A} \right)^{-1} \boldsymbol{A}^\top$.

Algorithm Outline: EM-PCA

Define

$$\begin{aligned} \boldsymbol{S} &= [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N] \in \mathbb{R}^{d \times N}, \\ \boldsymbol{X} &= [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}. \end{aligned}$$

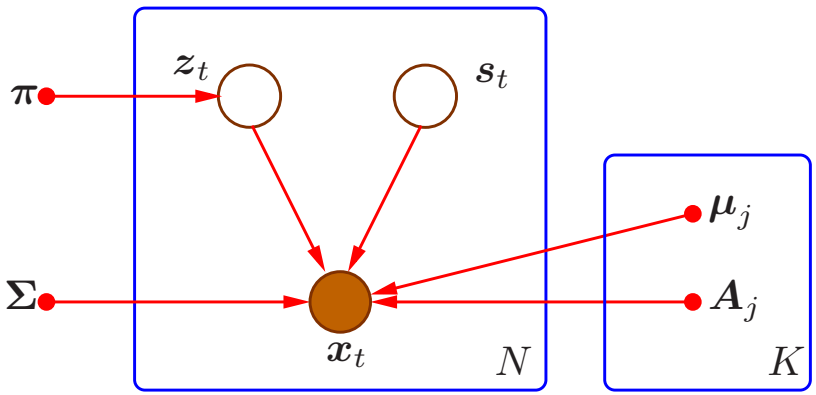
- E-step (LS projection):

$$\boldsymbol{S} = \left(\boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{X}.$$

- M-step (Wiener filter):

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{S}^\top \left(\boldsymbol{S}\boldsymbol{S}^\top\right)^{-1}.$$

Mixture of Factor Analyzers: Graphical Model



- Gaussian+Multinomial latent variables:

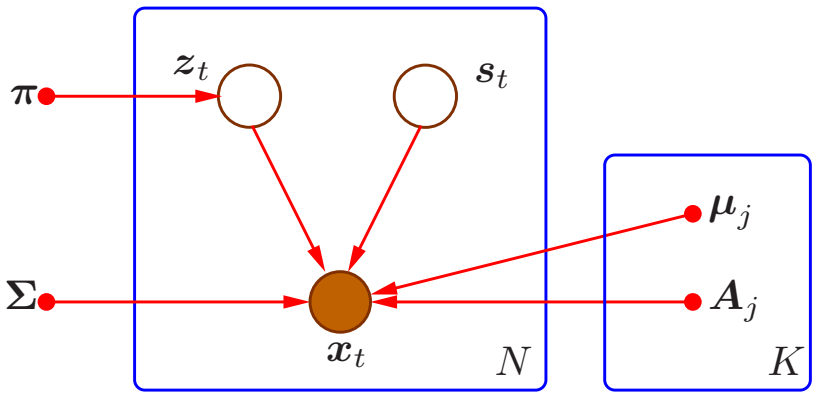
$$p(\boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{s}_t|0, \boldsymbol{I}),$$
$$p(\boldsymbol{z}_t) = \prod_{j=1}^{K} \pi_j^{z_{jt}}.$$

- The distribution over observed variables conditioned on latent variables is

$$p(\boldsymbol{x}_t|\boldsymbol{s}_t, \boldsymbol{z}_t) = \prod_{j=1}^{K} \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{A}_j\boldsymbol{s}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{z_{jt}}.$$

- Performs clustering and dimensionality reduction simultaneously.

Learning Mixture of Factor Analyzers



- Compute maximum likelihood estimates of parameters

$$\left\{\pi, \{A_j\}, \{\mu_j\}, \{\Sigma\}\right\}.$$

- Compute the posterior distribution over latent variables

$$p(s_t|x_t, z_{jt} = 1)$$

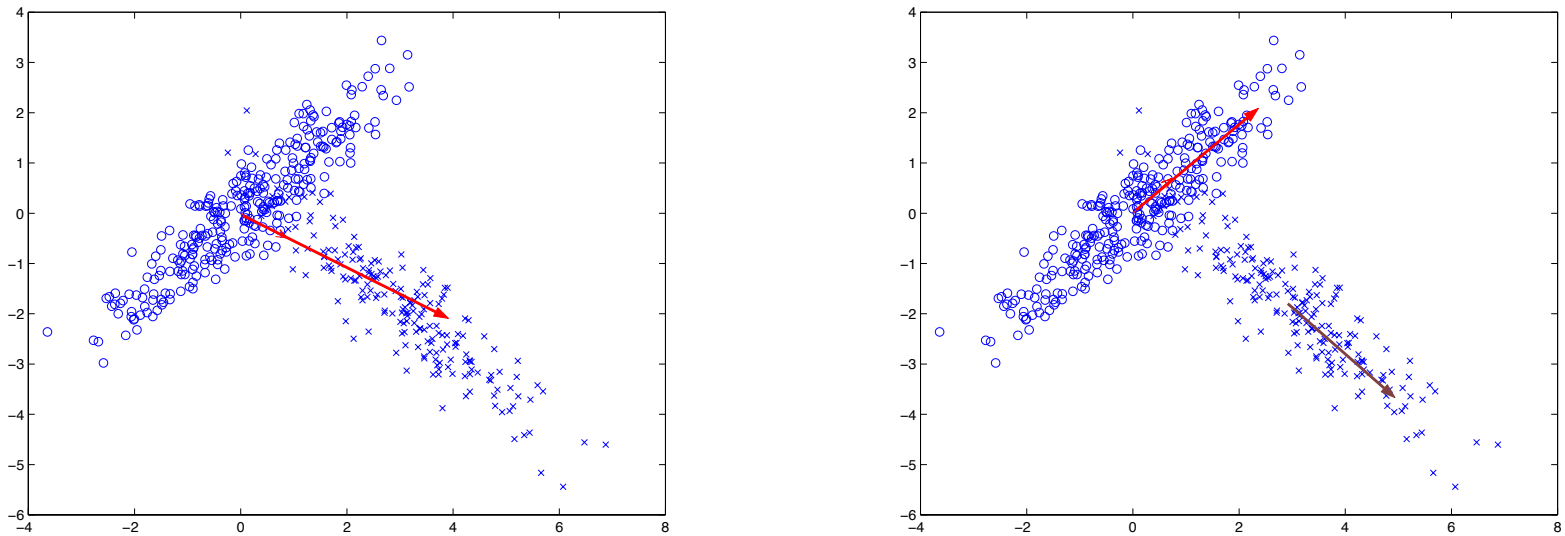
  which yields lower-dimensional projection in cluster j.

- We will do this again using expectation maximization.

Mixtures of Factor Analyzers (MFA)

- Clustering and dimensionality reduction are two fundamental problems in unsupervised learning.
  - The goal of clustering is to group data points by similarity between their features.
  - The goal of dimensionality reduction is to group (or compress) features that are highly correlated.
- Mixtures of Factor Analyzers = Dimensionality Reduction (FA) + Clustering (Gaussian Mixture Models).
- Might be the best method for local dimensionality reduction because it incorporate with probabilistic model.
- Good applications: character/face recognition

MFA: An Example



(a) (b)

Figure: (a) PCA; (b) Mixture of factor analyzers.

MFA: Complete-Data Log-Likelihood

- The complete-data log-likelihood is given by

$$\begin{aligned}\mathcal{L}_c &= \sum_{t=1}^{N} \log p(\boldsymbol{x}_t, \boldsymbol{s}_t, \boldsymbol{z}_t) \\ &= \sum_{t=1}^{N} \left[\log p(\boldsymbol{x}_t|\boldsymbol{s}_t, \boldsymbol{z}_t) + \log p(\boldsymbol{s}_t) + \log p(\boldsymbol{z}_t)\right],\end{aligned}$$

where ω_j represents $z_{jt} = 1$ and $\pi_j = p(\omega_j) = p(z_{jt} = 1)$, and

$$p(\boldsymbol{x}_t|\boldsymbol{s}_t, \boldsymbol{z}_t) = \prod_{j=1}^{K} \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{A}_j\boldsymbol{s}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{z_{jt}},$$

$$p(\boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{s}_t|0, \boldsymbol{I}), \qquad p(\boldsymbol{z}_t) = \prod_{j=1}^{K} \pi_j^{z_{jt}}.$$

- The posterior responsibility of the mixture j for generating data point $\boldsymbol{x}_t$, is defined by $\langle z_{jt}\rangle = R_{jt} = p(\omega_j|\boldsymbol{x}_t)$.
- The calculation of the expected complete-data log-likelihood, involves $\langle z_{jt}\boldsymbol{s}_t\rangle = R_{jt}\langle \boldsymbol{s}_t|\omega_j, \boldsymbol{x}_t\rangle$ and $\left\langle z_{jt}\boldsymbol{s}_t\boldsymbol{s}_t^\top\right\rangle = R_{jt}\left\langle \boldsymbol{s}_t\boldsymbol{s}_t^\top|\omega_j, \boldsymbol{x}_t\right\rangle$.

EM-MFA

- E-Step:

$$\begin{aligned} R_{jt} &= p(\omega_j|\boldsymbol{x}_t) \propto \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{A}_j\boldsymbol{A}_j^\top + \boldsymbol{\Sigma}), \\ \langle \boldsymbol{s}_t|\boldsymbol{x}_t, \omega_j\rangle &= \boldsymbol{\Phi}_j(\boldsymbol{x}_t - \boldsymbol{\mu}_j), \\ \langle \boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t, \omega_j\rangle &= \boldsymbol{I} - \boldsymbol{\Phi}_j\boldsymbol{A}_j + \boldsymbol{\Phi}_j(\boldsymbol{x}_t - \boldsymbol{\mu}_j)(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^\top\boldsymbol{\Phi}_j^\top. \end{aligned}$$

- M-Step:

$$\begin{aligned} \boldsymbol{A}_j &\leftarrow \left[\sum_t R_{jt}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)\langle \boldsymbol{s}_t|\boldsymbol{x}_t, \omega_j\rangle\right]\left[\sum_t R_{jt}\langle \boldsymbol{s}_t\boldsymbol{s}_t^\top|\boldsymbol{x}_t, \omega_j\rangle\right]^{-1}, \\ \boldsymbol{\mu}_j &\leftarrow \frac{\sum_t R_{jt}\left(\boldsymbol{x}_t - \boldsymbol{A}_j\langle \boldsymbol{s}_t|\boldsymbol{x}_t, \omega_j\rangle\right)}{\sum_t R_{jt}}, \\ \boldsymbol{\Sigma} &\leftarrow \frac{1}{N}\mathrm{diag}\left\{\sum_t\sum_j R_{jt}\left[(\boldsymbol{x}_t - \boldsymbol{\mu}_j)(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^\top - \boldsymbol{A}_j\langle \boldsymbol{s}_t|\boldsymbol{x}_t, \omega_j\rangle(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^\top\right]\right\}, \\ \pi_j &\leftarrow \frac{1}{N}\sum_t R_{jt}. \end{aligned}$$

Suggested Further Readings

1. C. Bishop (1999), "Latent variable models," In M. I. Jordan Ed., Learning in Graphical Models.
2. Z. Ghahramani and G. Hinton (1996), "The EM algorithm for mixtures of factor analyzers," University of Toronto Technical Report CRG-TR-96-1.
3. S. Roweis (1997), "EM algorithms for PCA and SPCA," NIPS-1997.
4. M. Tipping and C. Bishop (1999), "Probabilistic principal component analysis," Journal of the Royal Statistical Society, Series B.
5. M. Tipping and C. Bishop (1999), "Mixtures of probabilistic principal component analysers," Neural Computation.