Course 495: Advanced Statistical Machine Learning/Pattern Recognition

- Goal (Lecture): To present Probabilistic Principal Component Analysis (PPCA) using both Maximum Likelihood (ML) and Expectation Maximization (EM).

- Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding PPCA.

Materials

- Pattern Recognition & Machine Learning by C. Bishop Chapter 12
- **PPCA:** Tipping, Michael E., and Christopher M. Bishop. "Probabilistic principal component analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999): 611-622
- **PPCA:** Tipping, Michael E., and Christopher M. Bishop. "Mixtures of probabilistic principal component analyzers." *Neural computation* 11.2 (1999): 443-482.

An overview

PCA: Maximize the global variance

LDA: Minimize the class variance while maximizing the mean variance

LPP: Minimize the local variance

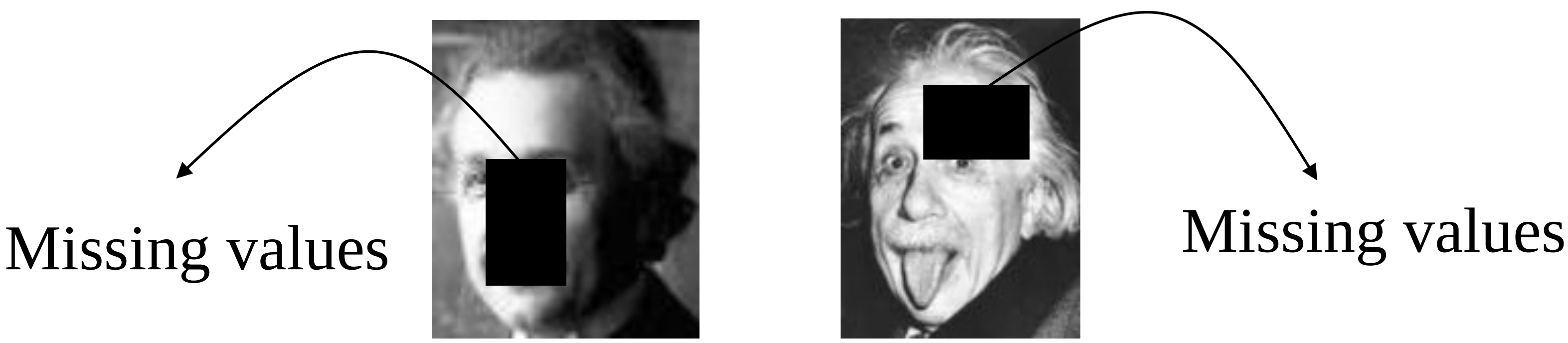
ICA: Maximize independence by maximizing non-Gaussianity

All are deterministic!!

Advantages of PPCA over PCA

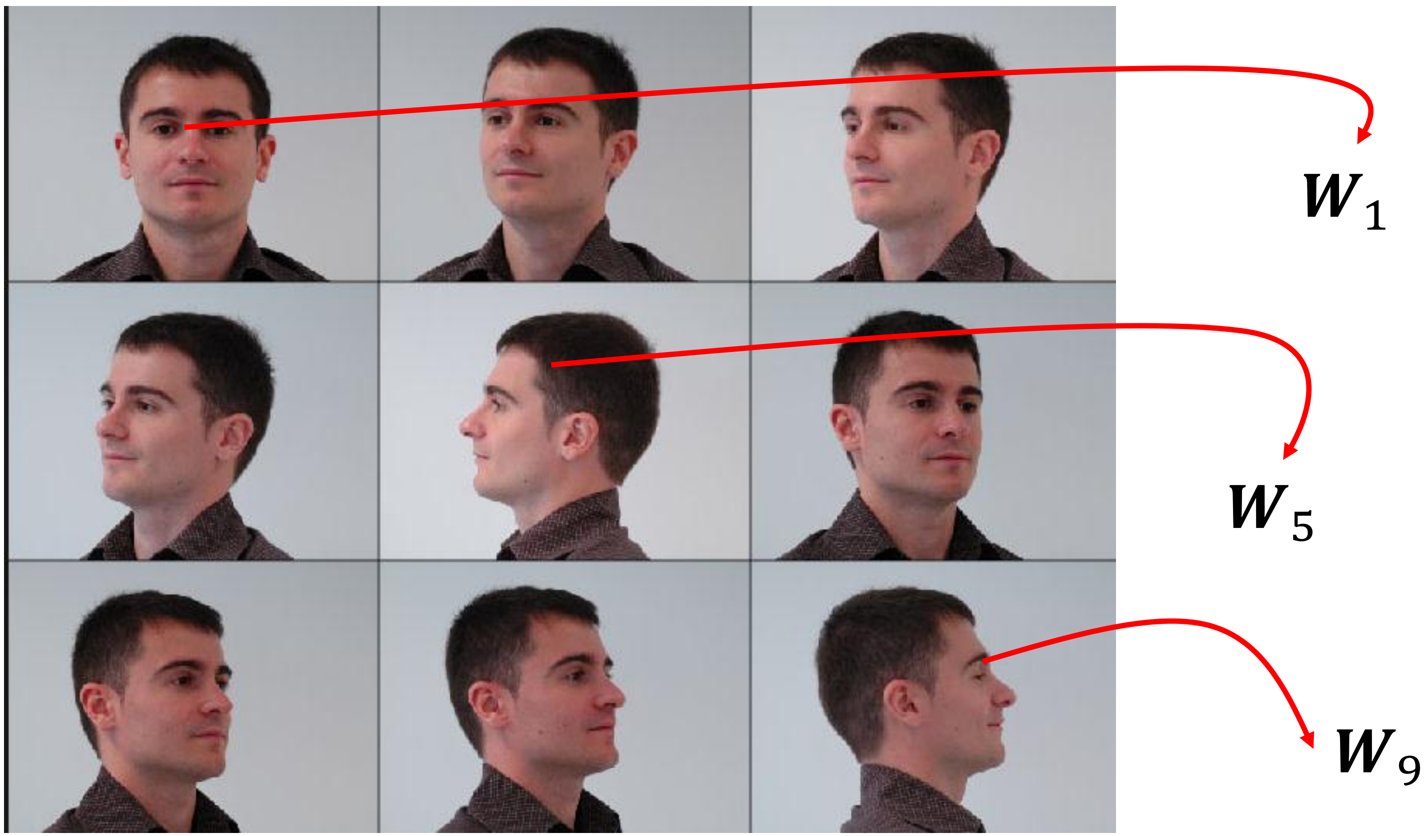
- An EM algorithm for PCA that is computationally efficient in situations where only a few leading eigenvectors are required and that avoids having to evaluate the data covariance matrix as an intermediate step.

- A combination of a probabilistic model and EM allows us to deal with missing values in the dataset.



Advantages of PPCA over PCA

- Mixtures of a probabilistic PCA models can be formulated in a principled way and trained using the EM algorithm.



Advantages of PPCA over PCA

- Probabilistic PCA forms the basis for a Bayesian treatment of PCA in which the dimensionality of the principal subspace can be found automatically from the data.

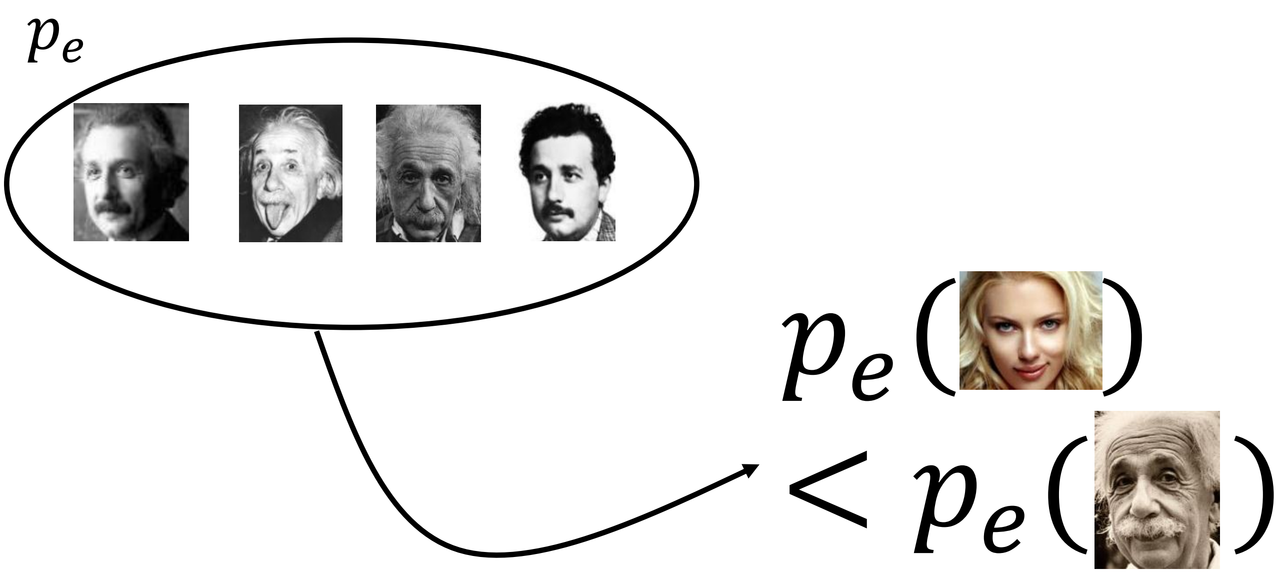
Priors on $\boldsymbol{W}$:Automatic Relevance Determination

(find the relevant components)

$$p(\boldsymbol{W}|\boldsymbol{a}) = \prod_{i=1}^{d} \left(\frac{a_i}{2\pi}\right)^{\frac{d}{2}} \exp\{-\frac{1}{2} a_i \boldsymbol{w}_i{}^T \boldsymbol{w}_i\}$$

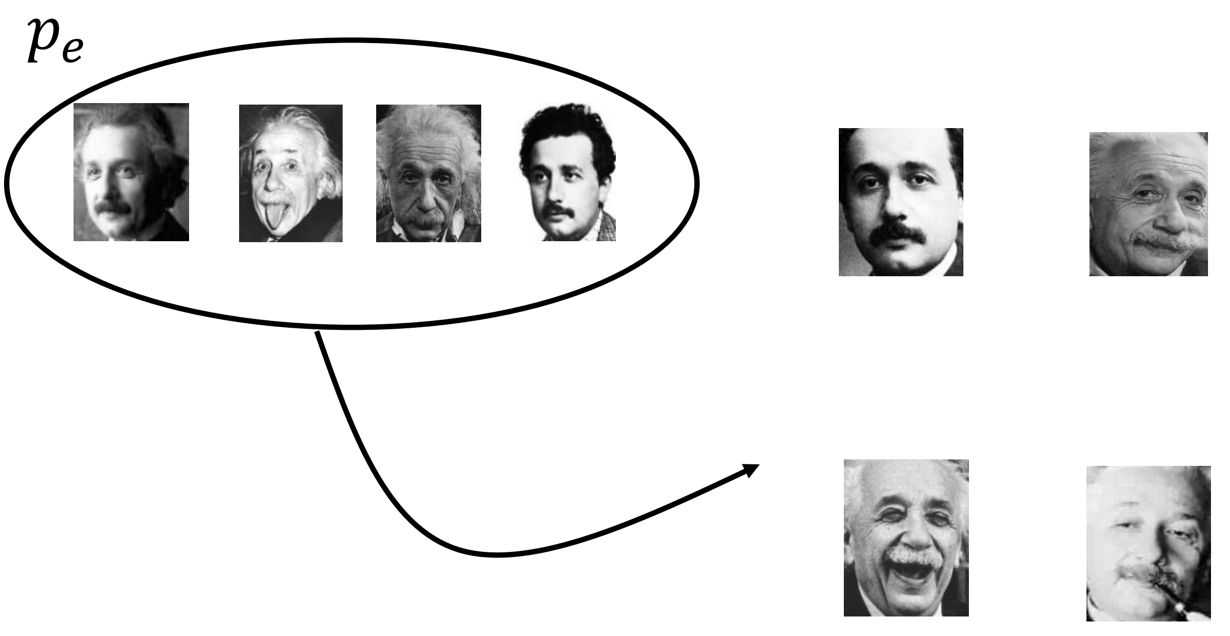
Advantages of PPCA over PCA

- Probabilistic PCA can be used to model class-conditional densities and hence be applied to classification problems.



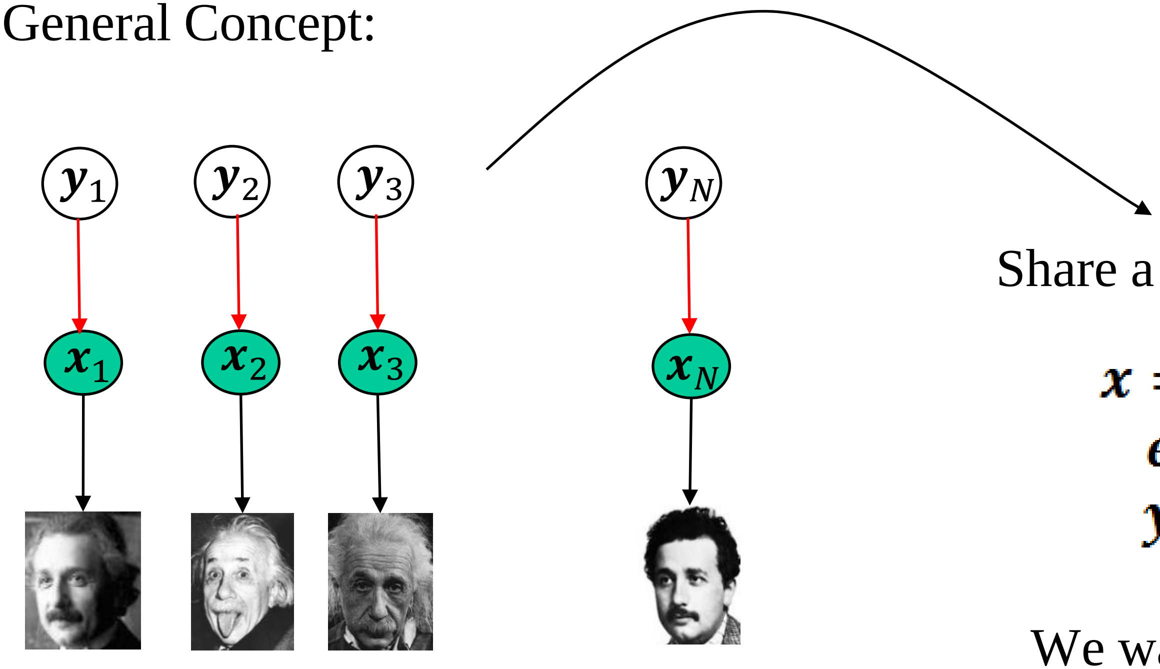
Advantages of PPCA over PCA

- The probabilistic PCA model can be run generatively to provide samples from the distribution.



Probabilistic Principal Component Analysis

General Concept:



Share a common linear structure

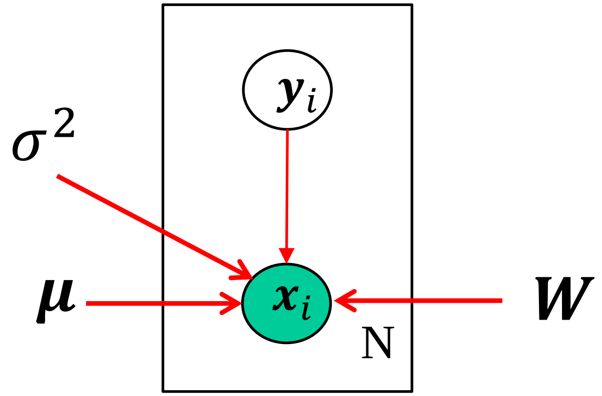
$$x = Wy + \mu + e$$
$$e \sim N(e|\mathbf{0}, \sigma^2 I)$$
$$y \sim N(y|\mathbf{0}, I)$$

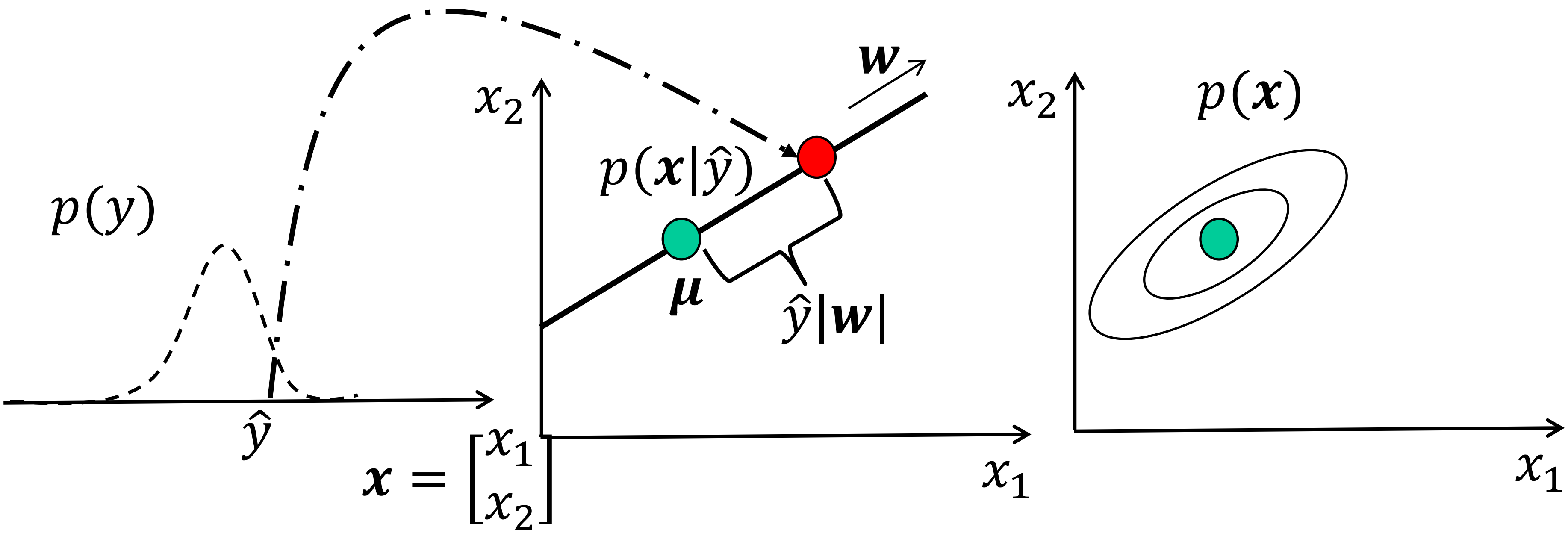
We want to find the parameters:

$$\theta = \{W, \mu, \sigma^2\}$$

Probabilistic Principal Component Analysis

Graphical Model:





ML-PPCA

$$p(\boldsymbol{x}_1,..,\boldsymbol{x}_{N,}\boldsymbol{y}_1,..,\boldsymbol{y}_N|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma) \prod_{i=1}^{\mathrm{N}} p(\boldsymbol{y}_i)$$

$$p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma) = \mathrm{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{y}_i+\boldsymbol{\mu},\sigma^2)$$

$$p(\boldsymbol{y}_i) = \mathrm{N}(\boldsymbol{y}_i|\boldsymbol{0},\boldsymbol{I})$$

$$\mathrm{p}(\boldsymbol{x}_1,..,\boldsymbol{x}_N\;|\theta) =$$

$$\int_{\boldsymbol{y}_i} \mathrm{p}(\boldsymbol{x}_1,..,\boldsymbol{x}_{N,}\boldsymbol{y}_1,..,\boldsymbol{y}_N|\theta)d\boldsymbol{y}_1 \dots d\boldsymbol{y}_N$$

$$= \int_{\boldsymbol{y}_i} \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma) \prod_{i=1}^{\mathrm{N}} p(\boldsymbol{y}_i)d\boldsymbol{y}_1 \dots d\boldsymbol{y}_N$$

ML-PPCA

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \,|\theta) = \int_{\boldsymbol{y}_i} \prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) \prod_{i=1}^{\mathrm{N}} p(\boldsymbol{y}_i) d\boldsymbol{y}_1 \ldots d\boldsymbol{y}_N$$

$$= \prod_{i=1}^{N} \int_{\boldsymbol{y}_i} p(\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{y}_i) d\boldsymbol{y}_i$$

$$p(\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{y}_i) = \frac{1}{\sqrt{(2\pi)^F \sigma^F}} e^{-\frac{1}{\sigma^2}(\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{y}_i)^T (\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{y}_i)}$$

$$\frac{1}{\sqrt{(2\pi)^d}} e^{-\boldsymbol{y}_i{}^T \boldsymbol{y}_i}$$

ML-PPCA

Complete the square:

$$\frac{1}{\sigma^2}\Big[(\underbrace{\boldsymbol{x}_i - \boldsymbol{\mu}}_{\bar{\boldsymbol{x}}_i} - \boldsymbol{W}\boldsymbol{y}_i)^T(\underbrace{\boldsymbol{x}_i - \boldsymbol{\mu}}_{\bar{\boldsymbol{x}}_i} - \boldsymbol{W}\boldsymbol{y}_i) + \sigma^2 \boldsymbol{y}_i{}^T \boldsymbol{y}_i\Big] =$$

$$\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{\mu}$$

$$= (\bar{\boldsymbol{x}}_i - \boldsymbol{W}\boldsymbol{y}_i)^T\ (\bar{\boldsymbol{x}}_i - \boldsymbol{W}\boldsymbol{y}_i) + \sigma^2 \boldsymbol{y}_i{}^T \boldsymbol{y}_i$$

$$= \bar{\boldsymbol{x}}_i{}^T\ \bar{\boldsymbol{x}}_i - 2\bar{\boldsymbol{x}}_i{}^T\boldsymbol{W}\boldsymbol{y}_i + \boldsymbol{y}_i{}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{y}_i + \sigma^2 \boldsymbol{y}_i{}^T \boldsymbol{y}_i$$

$$= \bar{\boldsymbol{x}}_i{}^T\ \bar{\boldsymbol{x}}_i - 2\bar{\boldsymbol{x}}_i{}^T\boldsymbol{W}\boldsymbol{y}_i + \boldsymbol{y}_i{}^T(\underbrace{\sigma^2\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{W}}_{\boldsymbol{M}})\boldsymbol{y}_i$$

ML-PPCA

$$= \bar{\boldsymbol{x}}_i{}^T\ \bar{\boldsymbol{x}}_i - 2\bar{\boldsymbol{x}}_i{}^T \boldsymbol{W}\boldsymbol{y}_i + \boldsymbol{y}_i{}^T \boldsymbol{M}\boldsymbol{y}_i$$

$$= \bar{\boldsymbol{x}}_i{}^T\ \bar{\boldsymbol{x}}_i - 2(\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)^T \boldsymbol{M}\boldsymbol{y}_i + \boldsymbol{y}_i{}^T \boldsymbol{M}\boldsymbol{y}_i + (\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)^T \boldsymbol{M} \boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i - (\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)^T \boldsymbol{M}\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i$$

$$= \bar{\boldsymbol{x}}_i{}^T\ (\boldsymbol{I} - \boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}^T)\bar{\boldsymbol{x}}_i + (\boldsymbol{y}_i - \boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)^{\boldsymbol{T}} \boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)$$

ML-PPCA

$$\int_{y_i} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma)p(\boldsymbol{y}_i)d\boldsymbol{y}_i \propto$$

$$\int_{y_i} e^{-\frac{1}{\sigma^2}\bar{\boldsymbol{x}}_i^T(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}^T)\bar{\boldsymbol{x}}_i-\frac{1}{\sigma^2}(\boldsymbol{y}_i-\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)^T\boldsymbol{M}(\boldsymbol{y}_i-\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)}d\boldsymbol{y}_i \propto$$

$$e^{-\frac{1}{\sigma^2}\bar{\boldsymbol{x}}_i^T(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}^T)\bar{\boldsymbol{x}}_i}\underbrace{\int_{y_i} e^{-\frac{1}{\sigma^2}(\boldsymbol{y}_i-\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)^T\boldsymbol{M}(\boldsymbol{y}_i-\boldsymbol{M}^{-1}\boldsymbol{W}^T\bar{\boldsymbol{x}}_i)}d\boldsymbol{y}_i}_{1}$$

$$= N(\boldsymbol{x}_i|\boldsymbol{\mu}, (\sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}^T)^{-1})$$

ML-PPCA

Let's have a look at:

$$\sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}^T = \sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}(\sigma^2\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$$

Woodbury formula:

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$$

Can you prove that if: $\boldsymbol{D} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\mathbf{I}$

then $\boldsymbol{D}^{-1} = \sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}(\sigma^2 I + \boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$

ML-PPCA

$$p(\boldsymbol{x}_i|\boldsymbol{W},\boldsymbol{\mu},\sigma^2) = \int_{y_i} p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma)p(\boldsymbol{y}_i)d\boldsymbol{y}_i = N(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{D})$$

$$\boldsymbol{D} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\mathbf{I}$$

Using also the above we can easily show that the posterior :

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma^2) = \frac{p(\boldsymbol{y}_i)}{p(\boldsymbol{x}_i|\boldsymbol{W},\boldsymbol{\mu},\sigma^2)}p(\boldsymbol{x}_i|\boldsymbol{y}_i,\boldsymbol{W},\boldsymbol{\mu},\sigma^2) =$$

$$= N(\boldsymbol{y}_i|\boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_i-\boldsymbol{\mu}),\sigma^2\boldsymbol{M}^{-1})$$

ML-PPCA

$$p(\boldsymbol{x}_1, .., \boldsymbol{x}_N \mid \theta) = \prod_{i=1}^{N} N(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{D})$$

$$\theta = \text{argmax}_\theta \ln p(\boldsymbol{x}_1, .., \boldsymbol{x}_N \mid \theta)$$

$$= \text{argmax}_\theta \sum_{i=1}^{N} \ln N(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{D})$$

$$= -\frac{NF}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{D}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{D}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

ML-PPCA

$$L(\boldsymbol{W}, \sigma^2, \boldsymbol{\mu}) = -\frac{NF}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{D}| - N\mathrm{tr}[\boldsymbol{D}^{-1}\boldsymbol{S}_t]$$

where $\boldsymbol{S}_t = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$

$$\frac{dL}{d\boldsymbol{\mu}} = \boldsymbol{0} \Rightarrow \boldsymbol{\mu} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\boldsymbol{x}_i$$

ML-PPCA

$$\frac{dL}{d\boldsymbol{W}} = N(\boldsymbol{D}^{-1}\boldsymbol{S}_t\boldsymbol{D}^{-1}\boldsymbol{W} - \boldsymbol{D}^{-1}\boldsymbol{W}) \Rightarrow \boldsymbol{S}_t\boldsymbol{D}^{-1}\boldsymbol{W} = \boldsymbol{W}$$

Three different solutions:

(1) $\boldsymbol{W} = \boldsymbol{0}$ which is the minimum of the likelihood

(2) $\boldsymbol{D} = \boldsymbol{S_t} \Rightarrow \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\mathbf{I} = \boldsymbol{S_t}$

Assume the eigen-decomposition $\boldsymbol{S_t} = \boldsymbol{U\Lambda U}^T$

$\boldsymbol{U}$ square matrix of eigenvectors $\boldsymbol{W} = \boldsymbol{U}(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R}$

$\boldsymbol{R}$ is a rotation matrix $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$

ML-PPCA

(3) $\boldsymbol{D} \neq \boldsymbol{S_t}$ and $\boldsymbol{W} \neq 0$ $\quad d < q = \text{rank}(\boldsymbol{S}_t)$

Assume the SVD of $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^{\boldsymbol{T}}$

$\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_d]$ Fxd matrix $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$

$$\boldsymbol{L} = \begin{bmatrix} l_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_d \end{bmatrix}$$

$$\boldsymbol{S}_t\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^{\boldsymbol{T}}$$

Let's study $\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{U}$

ML-PPCA

$$\boldsymbol{D}^{-1} = (\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \mathbf{I})^{-\mathbf{1}} \xRightarrow[\boldsymbol{W}=\boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T]{}$$

$$= (\boldsymbol{U}\boldsymbol{L}^2\boldsymbol{U}^T + \sigma^2 \mathbf{I})^{-\mathbf{1}}$$

Assume a set of bases $\boldsymbol{U}_{F-d}$ such that $\boldsymbol{U}_{F-d}{}^T\boldsymbol{U} = 0$ and $\boldsymbol{U}_{F-d}{}^T\boldsymbol{U}_{F-d} = \boldsymbol{I}$

$$= \left([\boldsymbol{U}\ \boldsymbol{U}_{F-d}] \begin{bmatrix} \boldsymbol{L}^2 & 0 \\ 0 & 0 \end{bmatrix} [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]^T + [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]\sigma^2\boldsymbol{I}[\boldsymbol{U}\ \boldsymbol{U}_{F-d}]^{\boldsymbol{T}} \right)^{-\mathbf{1}}$$

$$= [\boldsymbol{U}\ \boldsymbol{U}_{F-d}] \begin{bmatrix} \boldsymbol{L}^2 + \sigma^2\boldsymbol{I} & 0 \\ 0 & \sigma^2\boldsymbol{I} \end{bmatrix}^{-1} [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]^{\boldsymbol{T}}$$

$$= [\boldsymbol{U}\ \boldsymbol{U}_{F-d}] \begin{bmatrix} (\boldsymbol{L}^2 + \sigma^2\boldsymbol{I})^{-1} & 0 \\ 0 & \sigma^{-2}\boldsymbol{I} \end{bmatrix} [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]^{\boldsymbol{T}}$$

ML-PPCA

$$\boldsymbol{D}^{-1}\boldsymbol{U} = [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]\begin{bmatrix}(\boldsymbol{L}^2+\sigma^2\boldsymbol{I})^{-1} & 0 \\ 0 & \sigma^{-2}\mathrm{I}\end{bmatrix}[\boldsymbol{U}\ \boldsymbol{U}_{F-d}]^{\boldsymbol{T}}\boldsymbol{U}$$

$$= [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]\begin{bmatrix}(\boldsymbol{L}^2+\sigma^2\mathbf{I})^{-1} & 0 \\ 0 & \sigma^{-2}\boldsymbol{I}\end{bmatrix}[\boldsymbol{I}\ 0]^{\boldsymbol{T}}$$

$$= [\boldsymbol{U}\ \boldsymbol{U}_{F-d}]\begin{bmatrix}(\boldsymbol{L}^2+\sigma^2\boldsymbol{I})^{-1} \\ 0\end{bmatrix}$$

$$= \boldsymbol{U}(\boldsymbol{L}^2+\sigma^2\boldsymbol{I})^{-1}$$

ML-PPCA

$$\boldsymbol{S}_t \boldsymbol{D}^{-1} \boldsymbol{U} \boldsymbol{L} \boldsymbol{V}^T = \boldsymbol{U} \boldsymbol{L} \boldsymbol{V}^T$$

$$\boldsymbol{S}_{\boldsymbol{t}} \boldsymbol{U} (\boldsymbol{L}^2 + \sigma^2 \boldsymbol{I})^{-1} = \boldsymbol{U}$$

$$\boldsymbol{S}_{\boldsymbol{t}} \boldsymbol{U} = \boldsymbol{U} (\boldsymbol{L}^2 + \sigma^2 \boldsymbol{I})^{-1}$$

It means that

$$\boldsymbol{S}_t \boldsymbol{u}_{\boldsymbol{i}} = \left(l_i^{\ 2} + \sigma^2 \right) \boldsymbol{u}_i \qquad \boldsymbol{S}_t = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T$$

$\boldsymbol{u}_{\boldsymbol{i}}$ are eigenvectors of $\boldsymbol{S}_t$ and $\lambda_\iota = l_i^{\ 2} + \sigma^2 \Rightarrow l_\iota = \sqrt{\lambda_\iota - \sigma^2}$

Unfortunately $\boldsymbol{V}$ cannot be determined hence there is a rotation ambiguity.

ML-PPCA

Hence the optimum is given by (keeping d eigenvectors)

$$\boldsymbol{W}_d = \boldsymbol{U}_d(\boldsymbol{\Lambda}_d - \sigma^2\boldsymbol{I})\boldsymbol{V}^T$$

Having computed $\boldsymbol{W}$ we need to compute the optimum σ^2

$$L(\boldsymbol{W},\sigma^2,\boldsymbol{\mu}) = -\frac{NF}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{D}| - N\text{tr}[\boldsymbol{D}^{-1}\boldsymbol{S}_t]$$

$$= -\frac{NF}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\mathbf{I}\,| - N\text{tr}[(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}\,)^{-1}\boldsymbol{S}_t]$$

ML-PPCA

$$\boldsymbol{W}_d \boldsymbol{W}_d{}^T + \sigma^2 \mathbf{I} = [\boldsymbol{U}_d \boldsymbol{U}_{F-d}] \begin{bmatrix} \boldsymbol{\Lambda}_d - \sigma^2 \boldsymbol{I} & 0 \\ 0 & 0 \end{bmatrix} [\boldsymbol{U}_d \boldsymbol{U}_{F-d}]^T$$

$$+ [\boldsymbol{U}_d \boldsymbol{U}_{F-d}] \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} [\boldsymbol{U}_d \boldsymbol{U}_{F-d}]^T$$

$$= [\boldsymbol{U}_d \boldsymbol{U}_{F-d}] \begin{bmatrix} \boldsymbol{\Lambda}_d & 0 \\ 0 & \sigma^2 \boldsymbol{I} \end{bmatrix} [\boldsymbol{U}_d \boldsymbol{U}_{F-d}]^T$$

Hence $$|\boldsymbol{W}_d \boldsymbol{W}_d{}^T + \sigma^2 \boldsymbol{I}| = \prod_{i=1}^{d} \lambda_i \prod_{i=d+1}^{F} \sigma^2$$

$$\ln|\boldsymbol{W}_d \boldsymbol{W}_d{}^T + \sigma^2 \boldsymbol{I}| = (F-d)\ln\sigma^2 + \sum_{i=1}^{q} \ln\lambda_i$$

ML-PPCA

$$\boldsymbol{D}^{-1}\boldsymbol{S}_t = [\boldsymbol{U}_d\boldsymbol{U}_{F-d}]\begin{bmatrix}\boldsymbol{\Lambda}_d & 0\\ 0 & \sigma^2\boldsymbol{I}\end{bmatrix}^{-1}[\boldsymbol{U}_d\boldsymbol{U}_{F-d}]^T[\boldsymbol{U}_d\boldsymbol{U}_{F-d}]\begin{bmatrix}\boldsymbol{\Lambda}_d & 0 & 0\\ 0 & \boldsymbol{\Lambda}_{q-d} & 0\\ 0 & 0 & 0\end{bmatrix}[\boldsymbol{U}_d\boldsymbol{U}_{F-d}]^T$$

$$= [\boldsymbol{U}_d\boldsymbol{U}_{F-d}]\begin{bmatrix}\boldsymbol{I} & 0 & 0\\ 0 & \frac{1}{\sigma^2}\boldsymbol{\Lambda}_{q-d} & 0\\ 0 & 0 & 0\end{bmatrix}[\boldsymbol{U}_d\boldsymbol{U}_{F-d}]^T$$

$$\Rightarrow \mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{S}_t) = \frac{1}{\sigma^2}\sum_{i=d+1}^{q}\lambda_i + \mathrm{d}$$

ML-PPCA

$$L(\sigma^2) = -\frac{N}{2}\left\{F\ln 2\pi + \sum_{j=1}^{d} \ln \lambda_j + \frac{1}{\sigma^2}\sum_{j=d+1}^{q} \lambda_j + (F-d)\ln\sigma^2 + d\right\}$$

$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow -2\sigma^{-3}\sum_{j=d+1}^{q} \lambda_j + \frac{F-d}{\sigma} = 0 \Rightarrow \sigma^2 = \frac{1}{F-d}\sum_{j=d+1}^{q} \lambda_j$$

If I put the solution back

$$L(\sigma^2) = -\frac{N}{2}\left\{\sum_{j=1}^{d} \ln \lambda_j + (F-q)\ln\frac{1}{F-d}\sum_{j=d+1}^{F} \lambda_j + F\ln 2\pi + F\right\}$$

ML-PPCA

$$L(\sigma^2) = -\frac{N}{2}\{\underbrace{\sum_{j=1}^{d}\ln\ \lambda_j + \sum_{j=d}^{q}\ln\ \lambda_j}_{\ln|\boldsymbol{S}_t|} - \sum_{j=d}^{q}\ln\ \lambda_j + (F-d)\ln\frac{1}{F-d}\sum_{j=d+1}^{F}\lambda_j + F\ln 2\pi + F\}$$

$$\max\frac{N}{2}\left\{\frac{1}{F-d}\ln|\boldsymbol{S}_t| - \frac{1}{F-d}\sum_{\mathrm{j=d}}^{\mathrm{q}}\ln\lambda_j + \ln\left(\frac{1}{F-d}\sum_{\mathrm{j=d+1}}^{\mathrm{F}}\ln\lambda_j\right) + \mathrm{cost}\right\}$$

$$\Rightarrow \min\ln\left(\frac{1}{F-d}\sum_{\mathrm{j=d}}^{\mathrm{q}}\ln\lambda_j\right) - \frac{1}{F-d}\sum_{\mathrm{j=d}}^{\mathrm{q}}\ln\lambda_j$$

ML-PPCA

Jensen inequality $\ln\left(\frac{\sum_{i=1}^{n} r_{\mathrm{i}}}{n}\right) \geq \frac{1}{n}\sum_{i=1}^{n}\ln r_i$

$$\ln\left(\frac{1}{F-d}\sum_{j=d+1}^{q}\lambda_j\right) \geq \frac{1}{F-d}\sum_{j=d+1}^{q}\ln\lambda_j \ \text{ hence}$$

$$\Rightarrow \ln\left(\frac{1}{F-d}\sum_{j=d}^{q}\ln\lambda_j\right) - \frac{1}{F-d}\sum_{j=d}^{q}\ln\lambda_j \geq 0$$

Hence, the function is minimized when the discarded eigenvectors are the ones that correspond to the $q-d$ eigenvalues

ML-PPCA

Summarize:

$$\sigma^2 = \frac{1}{F-d}\sum_{j=d+1}^{q}\lambda_j$$

$$\boldsymbol{W_d} = \boldsymbol{U_d}(\boldsymbol{\Lambda_d} - \sigma^2\boldsymbol{I})\boldsymbol{V^T}$$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

We no longer have a projection but:

$$\boldsymbol{E}_{\boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i})}\{\boldsymbol{y_i}\} = \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_i - \boldsymbol{\mu})$$

and a reconstruction

$$\hat{\boldsymbol{x}}_i = \boldsymbol{W}\boldsymbol{E}_{\boldsymbol{p}(\boldsymbol{y_i}|\boldsymbol{x_i})}\{\boldsymbol{y_i}\} + \boldsymbol{\mu}$$

ML-PPCA

Hence

$$\lim_{\sigma^2 \to 0} \boldsymbol{W}_d = \boldsymbol{U}_d \boldsymbol{\Lambda}_d$$

$$\lim_{\sigma^2 \to 0} \boldsymbol{E}_{\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{x}_i)}\{\boldsymbol{y}_i\} = \boldsymbol{M}^{-1} {\boldsymbol{W}_d}^T (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= {\boldsymbol{\Lambda}_d}^{-1} \boldsymbol{U}_d \, (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$\lim_{\sigma^2 \to 0} \boldsymbol{M} = {\boldsymbol{W}_d}^T \boldsymbol{W}_d$$

which gives the whitened PCA

EM-PPCA

First step formulate the joint likelihood

$$p(\boldsymbol{X},\boldsymbol{Y}|\theta) = p(\boldsymbol{X},\boldsymbol{Y}|\theta)p(\boldsymbol{Y}) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{p}(\boldsymbol{x}_i|\boldsymbol{y}_i,\theta)\mathrm{p}(\boldsymbol{y}_i)$$

$$\ln p(\boldsymbol{X},\boldsymbol{Z}|\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \{\ln p(\boldsymbol{x}_i|\boldsymbol{y}_i,\theta) + \ln p(\boldsymbol{y}_i)\}$$

$$\ln p(\mathbf{x}_i|\mathbf{y}_i,\theta) = \ln \frac{1}{\sqrt{(2\pi)^{\mathrm{F}}(\sigma^2)^F}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{y}_i-\boldsymbol{\mu})^T(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{y}_i-\boldsymbol{\mu})}$$

$$= -\frac{\mathrm{F}}{2}\ln\left(\sqrt{2\pi}\sigma^2\right) - \frac{1}{2\sigma^2}(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{y}_i-\boldsymbol{\mu})^T(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{y}_i-\boldsymbol{\mu})$$

$$\ln \mathrm{p}(\boldsymbol{y}_i) = -\frac{1}{2}{\boldsymbol{y}_i}^T\boldsymbol{y}_i - \frac{\mathrm{D}}{2}\ln 2\pi$$

EM-PPCA

$$\ln \mathrm{p}(\boldsymbol{X}, \boldsymbol{Y}|\theta) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \{-\frac{\mathrm{F}}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu})^T(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i - \boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{y}_i{}^T\boldsymbol{y}_i - \frac{\mathrm{D}}{2}\ln 2\pi\}$$

I need now to optimize it with regards to $\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2$

We can't hence we need to take the expectations

EM-PPCA

But first we need to expand a bit:

$$\ln p(\boldsymbol{X}, \boldsymbol{Y}|\theta) = -\frac{NF}{2}\ln 2\pi\sigma^2 - \frac{ND}{2}\ln 2\pi - \sum_{i=1}^{N}\left\{\frac{1}{2\sigma^2}[(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu}) - 2(\mathbf{W}^T\mathbf{x}_i)^T\boldsymbol{y}_i + \mathbf{y}_i{}^T\mathbf{W}^T\mathbf{W}\boldsymbol{y}_i]\right\} + \frac{1}{2}\mathbf{y}_i{}^T\boldsymbol{y}_i$$

EM-PPCA

$$\mathbf{E}_{P(\boldsymbol{Y}|\boldsymbol{X})}\{\log \mathrm{P}(\boldsymbol{X},\boldsymbol{Y}|\theta)\} = -\frac{\mathrm{NF}}{2}\ln 2\pi\sigma^2 - \frac{\mathrm{ND}}{2}\ln 2\pi$$

$$-\sum_{i=1}^{N}\{\frac{1}{2\sigma^2}\left[(\mathbf{x}_i-\boldsymbol{\mu})^T(\mathbf{x}_i-\boldsymbol{\mu}) - \mathbf{2}(\mathbf{W}^T\mathbf{x}_i)^T\mathrm{E}\{\boldsymbol{y}_i\} + \mathrm{tr}[\mathrm{E}\{\boldsymbol{y}_i\mathbf{y}_i{}^T\}\mathbf{W}^T\mathbf{W}]\right.$$

EM-PPCA

$$E\{\mathbf{y}_i\} = \int \mathbf{y}_i \mathrm{p}(\mathbf{y}_i|\mathbf{x}_i, \theta)\mathrm{d}\mathbf{y}_n$$

$$= E\{N(\mathbf{y}_i|\boldsymbol{M}^{-1}\boldsymbol{W}^T(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1})\}$$

$$= \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$E\{(\mathbf{y}_i - E\{\mathbf{y}_i\})(\mathbf{y}_{\boldsymbol{i}} - E\{\mathbf{y}_i\})^T\} =$$

$$= E\{(\mathbf{y}_i\mathbf{y}_i{}^T)\} - E\{E\{\mathbf{y}_i\}\mathbf{y}_i\}^T - E\{\mathbf{y}_i E\{\mathbf{y}_i\}^T\} + E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^{\boldsymbol{T}}$$

$$= E\{\mathbf{y}_i\mathbf{y}_i{}^T\} - E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^T - E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^T + E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^{\boldsymbol{T}}$$

$$= E\{\mathbf{y}_i\mathbf{y}_i{}^T\} - E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^T = \sigma^2\mathbf{M}^{-1}$$

hence $E\{\mathbf{y}_i\mathbf{y}_i{}^T\} = \sigma^2\mathbf{M}^{-1} + E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^T$

EM-PPCA

Expectation Step:

Given $\boldsymbol{W}, \boldsymbol{\mu}, \sigma$:

$$E\{\mathbf{y}_i\} = \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$E\{\mathbf{y}_i{\mathbf{y}_i}^T\} = \sigma^2\mathbf{M}^{-1} + E\{\mathbf{y}_i\}E\{\mathbf{y}_i\}^T$$

EM-PPCA

Maximization step

$$\frac{\partial E\{\boldsymbol{\mu}\}}{\partial \boldsymbol{\mu}} = \mathbf{0} \Rightarrow \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\frac{\partial E\{\boldsymbol{W}\}}{\partial \boldsymbol{W}} = \mathbf{0} \Rightarrow \sum_{i=1}^{N}\left(-\frac{1}{\sigma^2}(\boldsymbol{x}_i - \boldsymbol{\mu})E\{\boldsymbol{y}_i\}^T + \frac{2}{2\sigma^2}\boldsymbol{W}E\{\boldsymbol{y}_i\boldsymbol{y}_i{}^T\}\right) = 0$$

$$\Rightarrow \boldsymbol{W} = \left[\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})E\{\boldsymbol{y}_i\}^T\right]\left[\sum_{i=1}^{N}E\{\mathbf{y}_i\boldsymbol{y}_i{}^T\}\right]^{-1}$$

EM-PPCA

$$\frac{\partial \mathbf{E}\{\sigma^2\}}{\partial \sigma} = \mathbf{0} \Rightarrow$$

$$\boldsymbol{\sigma^2}$$

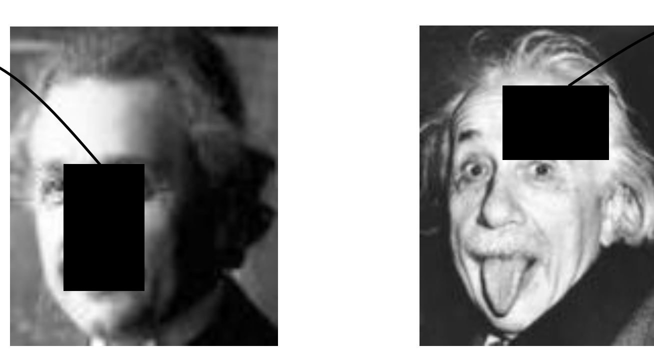
$$= \frac{1}{NF}\sum_{i=1}^{N}\{\left||\boldsymbol{x_i} - \boldsymbol{\mu}|\right|^{\mathbf{2}} - 2E\{\boldsymbol{y}_i\}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \mathrm{tr}[E\{\mathbf{y}_i\mathbf{y}_i{}^T\}\boldsymbol{W}^T\boldsymbol{W}]\}$$

Why EM-PPCA?

- A complexity of $O(NFd)$ which can be significant smaller than $O(NF^2)$ (computation of the covariance)

- A combination of a probabilistic model and EM allows us to deal with missing values in the dataset.

Missing values



Missing values

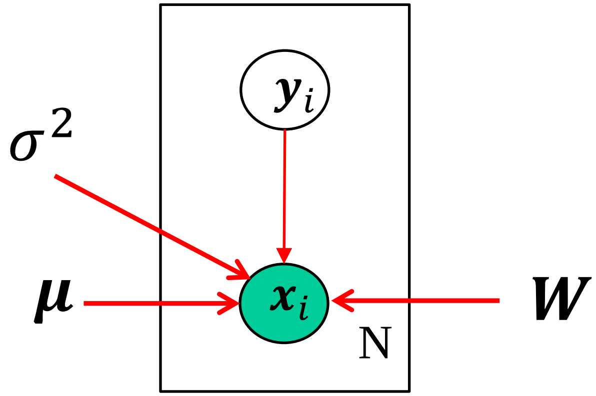
Why EM-PPCA?

A combination of a probabilistic model and EM allows us to deal with missing values in the dataset.

$$p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}_1, \boldsymbol{\mu}_1, \sigma)$$
$$\vdots$$
$$p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}_9, \boldsymbol{\mu}_9, \sigma)$$

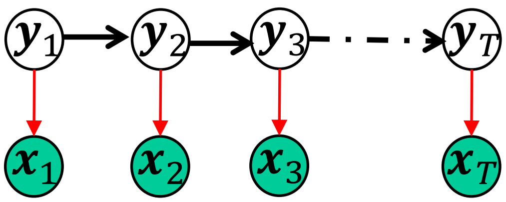
Summary

- We saw how to perform parameter estimation and inference using ML and EM in the following graphical model



What will we see next?

Dynamic data



EM in the following graphical models

Spatial data