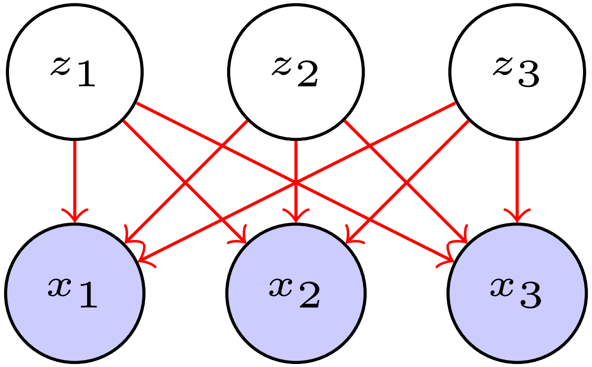
Chapter 13

Linear Factor Models

Linear Factor Models

- A probabilistic model $p(\boldsymbol{x}, \boldsymbol{z})$ with latent variables $\boldsymbol{z}$ that generates visible variables $\boldsymbol{x}$ by adding noise $\boldsymbol{\epsilon}$ to an affine function of $\boldsymbol{z}$



- In symbols, we have

$$\boldsymbol{z} \sim p(\boldsymbol{z}) \quad \text{Gaussian}$$

$$\boldsymbol{x} = \underbrace{\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}}_{\text{Affine}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}} \quad ?$$

? ?

- The latent variables $\boldsymbol{z}$ capture the dependencies between the observed data $\boldsymbol{x}$ and are known as explanatory factors

- Generally, $p(\boldsymbol{z})$ is assumed to be factorial, i.e.,

$$p(\boldsymbol{z}) = \prod_i p(z_i)$$

and the noise $\boldsymbol{\epsilon}$ is a Gaussian and is independent of $\boldsymbol{z}$

$$p(\boldsymbol{\epsilon}) \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, \sigma^2 \boldsymbol{I})$$

- It then follows that the conditional probability $p(\boldsymbol{x}|\boldsymbol{z})$ is given by

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$$

$p(x) = \int p(x, z) dz$

- With these, we have a complete probabilistic model

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}),$$

assuming all model parameters $\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2$ are known

- In principle, we can
  - Do any probabilistic inference, e.g., to predict $\boldsymbol{z}$ based on $\boldsymbol{x}$

$$p(\boldsymbol{z}|\boldsymbol{x}) \propto p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$$

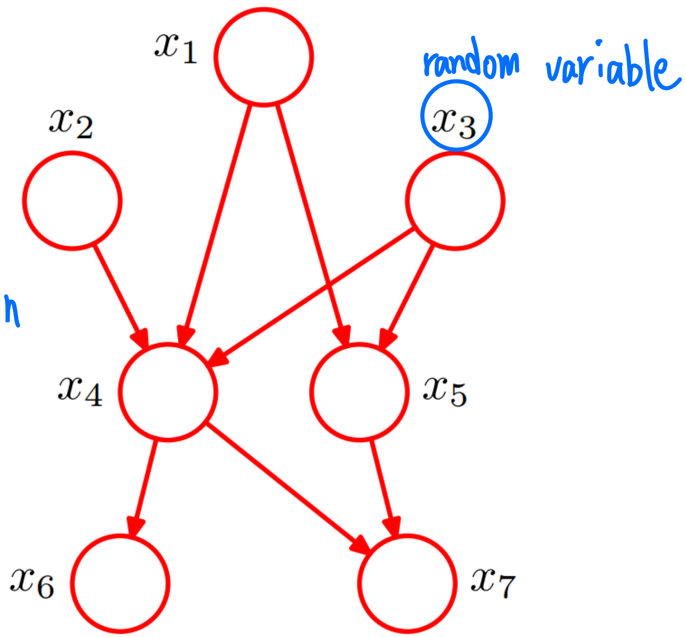
  - Generate $\boldsymbol{x}$ by first sampling $\boldsymbol{z}$ and then using $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$
  - etc.

Graphical Models 101

- To represent the factorization of a probability distribution using a graph

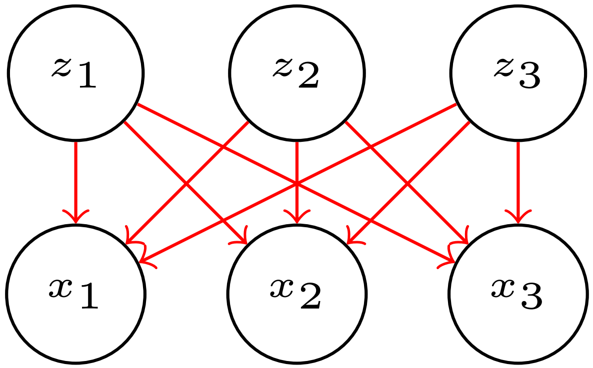
$p(x_5 | x_1, x_3)$ graphical model

$= p(x_5 | x_1, x_2, x_3, x_4)$ chain rule



$$p(x_1, x_2, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$
$$p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

- Applying the same principle, we have for the following graphical model



$$
\begin{aligned}
p(x_1, x_2, x_3, z_1, z_2, z_3) = {} & p(z_1)p(z_2)p(z_3) \\
& p(x_1|z_1, z_2, z_3)p(x_2|z_1, z_2, z_3)p(x_3|z_1, z_2, z_3)
\end{aligned}
$$

- This implies that x_1, x_2, x_3 are conditionally independent given z_1, z_2, z_3, a property that can be obtained by examining the graph

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{z}) &= \frac{p(x_1, x_2, x_3, z_1, z_2, z_3)}{p(z_1)p(z_2)p(z_3)} \\
&= p(x_1|z_1, z_2, z_3)p(x_2|z_1, z_2, z_3)p(x_3|z_1, z_2, z_3)
\end{aligned}
$$

Probabilistic Principle Component Analysis (PCA)

- Probabilistic PCA is one example of linear factor models with

prior :
$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$$
$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$$

- The conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$ suggests

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon}$ being independent of $\boldsymbol{z}$ and following a Gaussian distribution

$$p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

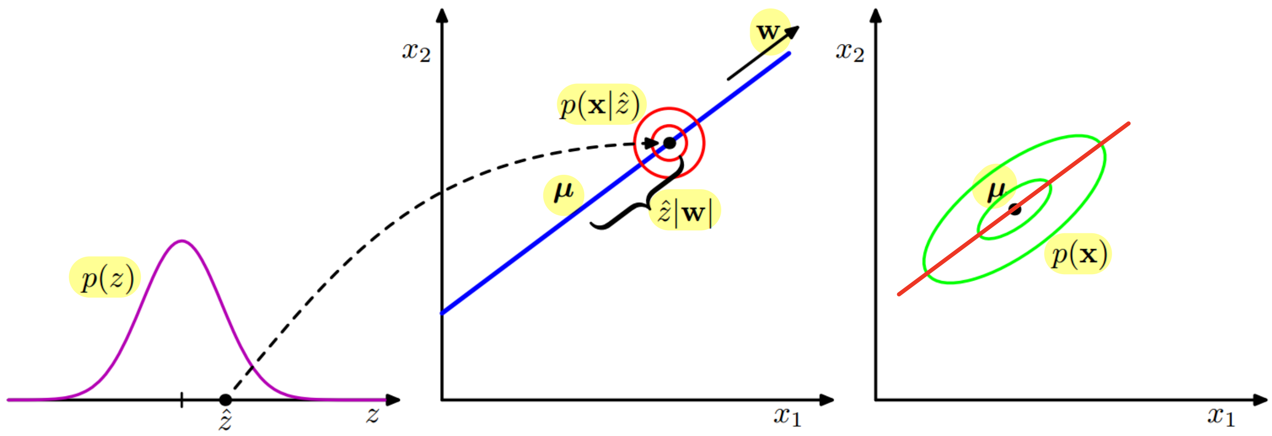
- It is assumed that the observed variable $\boldsymbol{x}$ is D-dimensional, and the latent variable $\boldsymbol{z}$ is M-dimensional

$D \gg M$

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} z + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$

const.

- Example: 2-D observed variable $\boldsymbol{x}$ + 1-D latent variable z



- By noting that $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$, and $\boldsymbol{z}, \boldsymbol{\epsilon}$ are independent, one can deduce the marginal distribution $p(\boldsymbol{x})$ is a Gaussian

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{C}),$$

$X = \begin{pmatrix} W & I \end{pmatrix} \begin{pmatrix} z \\ \varepsilon \end{pmatrix} = Wz + \varepsilon$

whose covariance matrix $\boldsymbol{C}$ is given by

$$E((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T) = E((\boldsymbol{W}\boldsymbol{z}+\boldsymbol{\epsilon})(\boldsymbol{W}\boldsymbol{z}+\boldsymbol{\epsilon})^T)$$
$$= E(\boldsymbol{W}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{W}^T) + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)$$

if this is Gaussian
then x is too.

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

- Remarks
  - The resulting Gaussian distribution $p(\boldsymbol{x})$ is governed by $\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2$, which generally have a smaller parameter count $(D + DM + 1)$ than direct specification $(D + \frac{D(D+1)}{2})$ of a general Gaussian
  - Applying any unitary rotation $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$ to the latent space $\tilde{\boldsymbol{z}} = \boldsymbol{R}\boldsymbol{z}$ does not change $p(\boldsymbol{x})$; as can be seen,

$$E(\boldsymbol{W}\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^T\boldsymbol{W}^T) = E(\underbrace{\boldsymbol{W}\boldsymbol{R}}_{\tilde{\boldsymbol{W}}}\boldsymbol{z}\boldsymbol{z}^T\underbrace{\boldsymbol{R}^T\boldsymbol{W}^T}_{\tilde{\boldsymbol{W}}^T}) = \boldsymbol{W}\boldsymbol{W}^T$$

  - This suggests that there are a family of $\tilde{\boldsymbol{W}}$ that lead to the same $p(\boldsymbol{x})$, and we may need to additionally specify $\boldsymbol{R}$ in order to identify the true $\boldsymbol{W}$

- The posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$ can be evaluated as a Gaussian

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{\mu})), \sigma^2\boldsymbol{M}^{-1})$$

where

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

- This is solved straightforwardly by observing that

$$p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$$

has a quadratic form in $\boldsymbol{z}$ in the resulting exponent; that is,

$$\begin{aligned}
\underbrace{p(\boldsymbol{z})}_{\text{Gass.}}\underbrace{p(\boldsymbol{x}|\boldsymbol{z})}_{\text{Gass.}} &= c\exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}\right) \\
&\propto c'\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right) \\
&= \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= p(\boldsymbol{z}|\boldsymbol{x})
\end{aligned}$$

Maximum Likelihood PCA

- To determine the model parameters $\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2$, the maximum likelihood (ML) principle can be applied to maximize

$$
\begin{aligned}
&\log p(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \\
&= \sum_n^N \log p(\boldsymbol{x}_n; \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \\
&= \left(-\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\boldsymbol{C}| - \frac{1}{2} \sum_{n=1}^N (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right)
\end{aligned}
$$

- Maximizing w.r.t. $\boldsymbol{u}$ is easy and leads to the sample mean

$$
\boldsymbol{u}_{\mathsf{ML}} = \bar{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^N \boldsymbol{x}_n
$$

maximize $z^T S z \Rightarrow \frac{1}{N} \sum z^T x_n x_n^T z$

s.t. $z^T z = 1$

- However, maximizing w.r.t. $\boldsymbol{W}$ and σ^2 needs some work, their closed-form solutions being

principal component

$$\begin{cases} \boldsymbol{W}_{\mathsf{ML}} = \boldsymbol{U}(\boldsymbol{L} - \sigma^2_{\mathsf{ML}}\boldsymbol{I})^{1/2}\boldsymbol{R} \\ \sigma^2_{\mathsf{ML}} = \frac{1}{D-M} \sum_{i=M+1}^{D} \lambda_i \end{cases}$$

$\mathcal{L}(z, \lambda) = z^T S z - \lambda(z^T z - 1)$

$\nabla_z \mathcal{L} = 2Sz - \lambda 2z = 0$

$\Rightarrow Sz = \lambda z$

$\Rightarrow z$ is eigenvector of S

  - $\boldsymbol{U}$ is a $D \times M$ matrix whose columns are given by the eigenvectors of the sample covariance matrix $\boldsymbol{S}$ that correspond to the largest M eigenvalues

$C = E[(x - \mu_x)(x - \mu_x)^T]$

$$\boldsymbol{S} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T$$

  - $\boldsymbol{L}$ is an $M \times M$ diagonal matrix whose elements are the corresponding eigenvalues λ_i
  - $\boldsymbol{R}$ is an arbitrary $M \times M$ unitary matrix (assumed to be $\boldsymbol{I}$ for

convenience)

- To summarize, we have a data model

$$p(\boldsymbol{z}) = \mathcal{N}(z; \boldsymbol{0}, \boldsymbol{I})$$
$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}_{\mathsf{ML}}\boldsymbol{z} + \boldsymbol{\mu}_{\mathsf{ML}}, \sigma^2_{\mathsf{ML}}\boldsymbol{I})$$

which gives

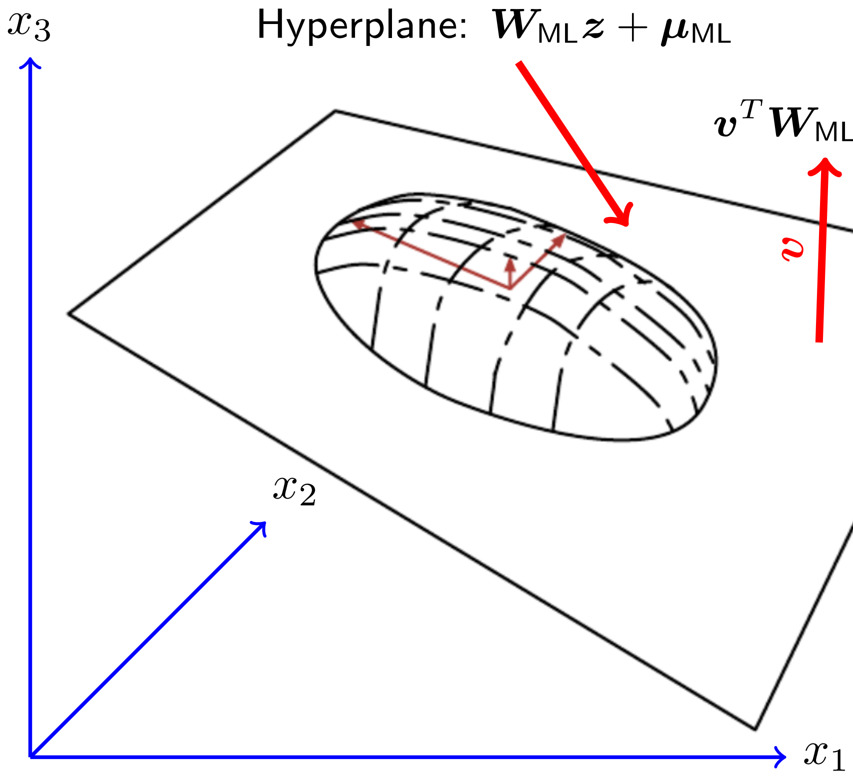
$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\mathsf{ML}}, \boldsymbol{C}_{\mathsf{ML}}) \text{ with } \boldsymbol{C}_{\mathsf{ML}} = \boldsymbol{W}_{\mathsf{ML}}\boldsymbol{W}^T_{\mathsf{ML}} + \sigma^2_{\mathsf{ML}}\boldsymbol{I}$$

- Observations
  - Along the principle axes $\boldsymbol{v} = \boldsymbol{U}_{:,i}$, the model correctly captures the data variance

$$E[(\boldsymbol{v}^T(\boldsymbol{x} - \boldsymbol{u}_{\mathsf{ML}}))^2] = \boldsymbol{v}^T\boldsymbol{C}_{\mathsf{ML}}\boldsymbol{v} = \lambda_i$$

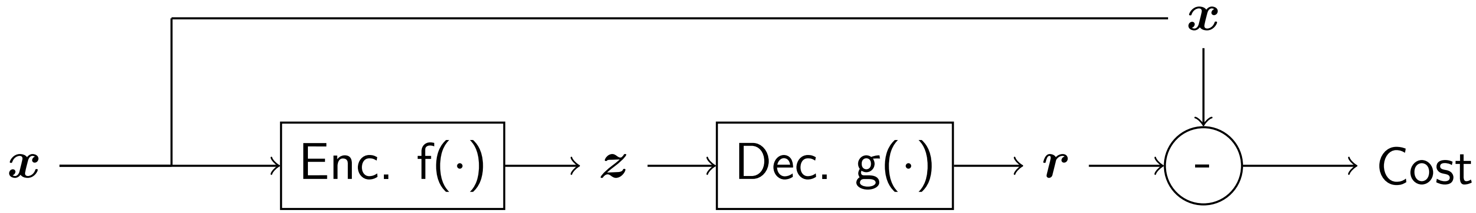
  - Along the axes $\boldsymbol{v}$ orthogonal to the principle subspace, i.e. $\boldsymbol{v}^T\boldsymbol{U} = \boldsymbol{0}$, the model predicts a variance that is the average of the discarded eigenvalues

$$E[(\boldsymbol{v}^T(\boldsymbol{x} - \boldsymbol{u}_{\mathsf{ML}}))^2] = \boldsymbol{v}^T \boldsymbol{C}_{\mathsf{ML}} \boldsymbol{v} = \sigma^2_{\mathsf{ML}}$$



Standard PCA

- Model setting (modified by introducing an affine decoder/encoder)



  - Input: $\boldsymbol{x} \in \mathbb{R}^D$
  - Representation: $\boldsymbol{z} \in \mathbb{R}^M$
  - Decoder: $g(\boldsymbol{z}) = \underbrace{\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}}$ with $\boldsymbol{W}$ having orthonormal columns
  - Cost: $\|\boldsymbol{x} - g(\boldsymbol{z})\|_2^2$ $\arg\min_{z} \|x - Wz - \mu\|_2^2$
  - Optimal encoder (when Cost minimized): $\boldsymbol{z} = f(\boldsymbol{x}) = \underbrace{\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu})}$

- To determine $\boldsymbol{\mu}$, we minimize the reconstruction error w.r.t. $\boldsymbol{\mu}$

$$\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{\mu}\|_2^2 = \sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{W}^T(\boldsymbol{x}_n - \boldsymbol{\mu}) - \boldsymbol{\mu}\|_2^2$$

$$\text{s.t. } \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I},$$

which gives

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n + \mathcal{C}(\boldsymbol{W}) = \bar{\boldsymbol{x}} + \mathcal{C}(\boldsymbol{W}),$$

where $\bar{\boldsymbol{x}}$ is sample mean and $\mathcal{C}(\boldsymbol{W})$ denotes the column space of $\boldsymbol{W}$.

- To determine $\boldsymbol{W}$, we minimize w.r.t. $\boldsymbol{W}$ the same objective yet expressed in the form used in Chapter 5

$$\arg\min_{\boldsymbol{W}} \|\tilde{\boldsymbol{X}}^{(\text{train})} - \tilde{\boldsymbol{X}}^{(\text{train})}\boldsymbol{W}\boldsymbol{W}^T\|_F^2, \text{ s.t. } \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$$

where

$$\tilde{\boldsymbol{X}}^{(\text{train})} = \begin{bmatrix} \boldsymbol{x}_1^{(\text{train})T} \\ \boldsymbol{x}_2^{(\text{train})T} \\ \vdots \\ \boldsymbol{x}_N^{(\text{train})T} \end{bmatrix} - \mathbf{1}\bar{\boldsymbol{x}}^T$$

with $\mathbf{1}$ denoting a column vector of 1's

- This allows us to follow the same line of derivations to conclude that the optimal $\boldsymbol{W}$ has its columns composed of the eigenvectors of the (scaled) sample covariance matrix $\tilde{\boldsymbol{X}}^{(\text{train})}\tilde{\boldsymbol{X}}^{(\text{train})T}$ that correspond to the largest M eigenvalues

$$\tilde{\boldsymbol{X}}^{(\text{train})}\tilde{\boldsymbol{X}}^{(\text{train})T} = \sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T$$

Standard PCA vs. Probabilistic PCA

- Standard PCA: Deterministic encoder/decoder

$$\text{Encoder: } \boldsymbol{z} = \boldsymbol{W}^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$$

$$\text{Decoder: } \boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \bar{\boldsymbol{x}}$$

- Probabilistic PCA: Stochastic encoder/decoder

$$\text{Encoder: } p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \bar{\boldsymbol{x}}), \sigma^2\boldsymbol{M}^{-1})$$

$$\text{Decoder: } p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \bar{\boldsymbol{x}}, \sigma^2\boldsymbol{I})$$

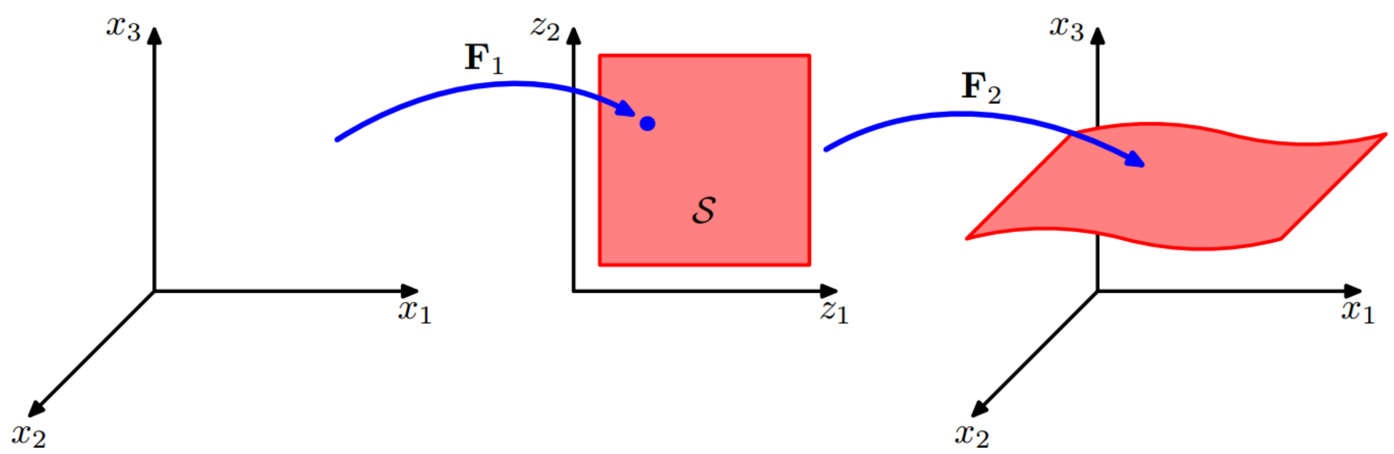
  where

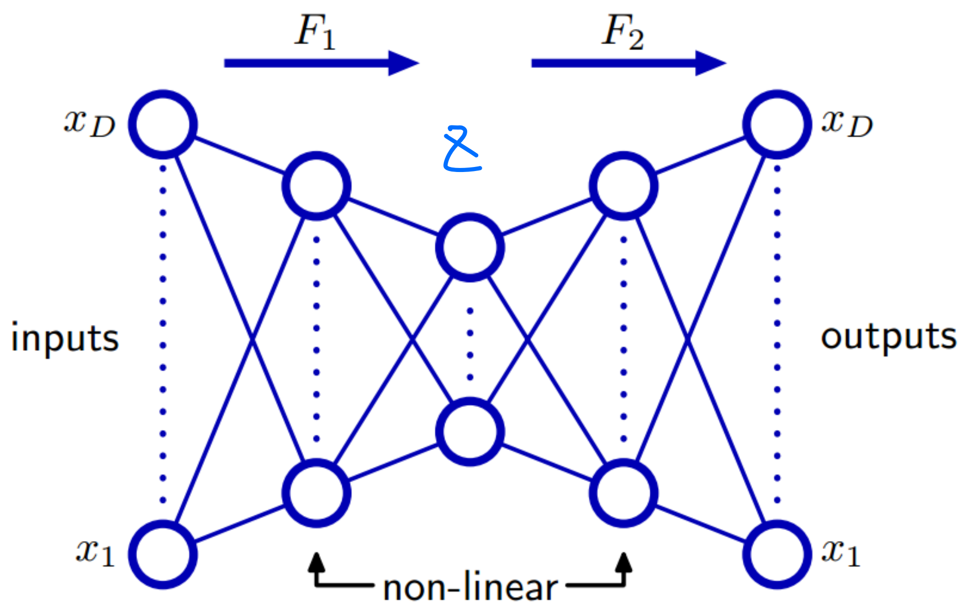
$$\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}$$

- When $\sigma^2 \rightarrow 0$, the standard PCA can be recovered from the probabilistic PCA

Manifold Interpretation of PCA

- Linear factor models, such as PCA, can be interpreted as learning a low-dimensional manifold
- Manifold in the present context is defined loosely to be a connected set of points with a small number of degrees of freedom, or dimensions, within a high-dimensional space
- Probabilistic PCA learns a pancake-shaped manifold of high probability
- Standard PCA learns a hyperplane specified by $\boldsymbol{W}\boldsymbol{z} + \bar{\boldsymbol{x}}$
- The idea of dimension reduction can be extended to incorporate neural networks to learn a general, non-linear manifold


x_3
x_1
x_2
$\mathbf{F}_1$
z_2
z_1
$\mathcal{S}$
$\mathbf{F}_2$
x_3
x_1
x_2


F_1
F_2
z
x_D
x_D
inputs
outputs
x_1
x_1
non-linear

The Expectation Maximization (EM) Algorithm

- A general technique for finding maximum likelihood (ML) solutions

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}; \boldsymbol{\theta})$$

for probabilistic models having latent variables $\boldsymbol{Z}$

$$p(\boldsymbol{X}; \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})$$

- Procedure

1. Choose an initial setting $\boldsymbol{\theta}^{\text{old}}$
2. **(E step)** Compute the expectation of the complete log-likelihood w.r.t. $\boldsymbol{Z}$ using the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\theta}^{\text{old}})$

$$E_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\theta}^{\text{old}})} \log p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})$$

3. **(M step)** Maximize the result w.r.t. $\boldsymbol{\theta}$ to give a new estimate $\boldsymbol{\theta}^{\text{new}}$

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} E_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}^{\text{old}})} \log p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})$$

4. Update $\boldsymbol{\theta}^{\text{old}}$ and repeat Steps 2-4 until convergence

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

- The EM algorithm is applicable when optimizing $p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})$ is easier than direct optimization of $p(\boldsymbol{X}; \boldsymbol{\theta})$

$$p(X,Z) = p(X)\,p(Z|X) \Rightarrow p(X) = \frac{p(X,Z)}{p(Z|X)}$$

- To see how the EM works, the chain rule of probability suggests

$$\log p(\boldsymbol{X};\boldsymbol{\theta}) = \log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}) - \log p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta})$$

- We next introduce an arbitrary distribution $q(\boldsymbol{Z})$ on both sides and integrate over $\boldsymbol{Z}$

$$
\begin{aligned}
\log p(X;\theta) = & \int q(\boldsymbol{Z}) \log p(\boldsymbol{X};\boldsymbol{\theta}) d\boldsymbol{Z} \\
= & \int q(\boldsymbol{Z}) \overset{1}{\log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})} d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \overset{2}{\log p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta})} d\boldsymbol{Z} \\
= & \underbrace{\int q(\boldsymbol{Z}) \overset{1}{\log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})} d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) d\boldsymbol{Z}}_{\mathcal{L}(X,q,\theta)} \\
& + \underbrace{\int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}) d\boldsymbol{Z}}_{\overset{2}{}}
\end{aligned}
$$

to arrive at $KL(q(Z)\,||\,p(Z|X;\theta))$

$$\log p(\boldsymbol{X};\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{X}, q, \boldsymbol{\theta}) + \mathsf{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}))$$

where

$$\mathcal{L}(\boldsymbol{X}, q, \boldsymbol{\theta}) = \int q(\boldsymbol{Z}) \log p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) d\boldsymbol{Z}$$

$$\mathsf{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta})) = \int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta})} d\boldsymbol{Z}$$

- Since the KL divergence is non-negative, $\mathsf{KL}(q||p) \geq 0$, it follows that

$$\log p(\boldsymbol{X};\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{X}, q, \boldsymbol{\theta})$$

  with equality if and only if

$$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta})$$

- In other words, $\mathcal{L}(\boldsymbol{X}, q, \boldsymbol{\theta})$ is a lower bound on $\log p(\boldsymbol{X};\boldsymbol{\theta})$

- Now, by choosing deliberately

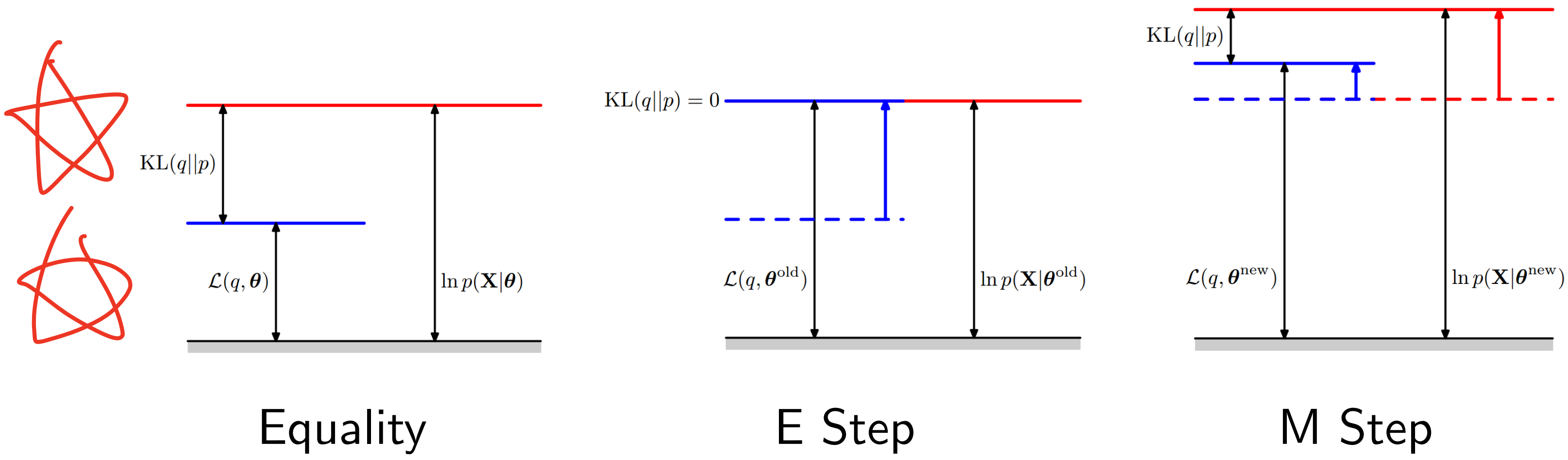
$$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}^{\mathsf{old}}),$$

we have

$$\begin{aligned}
\log p(\boldsymbol{X};\boldsymbol{\theta}^{\mathsf{new}}) &= \underbrace{\int q(\boldsymbol{Z}) \log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}^{\mathsf{new}}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) d\boldsymbol{Z}}_{(1)} \\
&+ \underbrace{\int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}^{\mathsf{new}})} d\boldsymbol{Z}}_{\geq 0} \\
&\geq \underbrace{\int q(\boldsymbol{Z}) \log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}^{\mathsf{old}}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z}) d\boldsymbol{Z}}_{(1')} \\
&+ \underbrace{\int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}^{\mathsf{old}})} d\boldsymbol{Z}}_{=0} = \log p(\boldsymbol{X};\boldsymbol{\theta}^{\mathsf{old}})
\end{aligned}$$

where $(1) \geq (1')$ is due to the M step

- The increase in $\log p(\boldsymbol{X};\boldsymbol{\theta})$ is at least as much as $\mathcal{L}(\boldsymbol{X}, q, \boldsymbol{\theta})$



EM for Probabilistic PCA

- The complete log-likelihood $\log p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$ is given by

$$\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{z}_n) + \log p(\boldsymbol{z}_n)$$

- In the E step, we take expectation of the log-likelihood w.r.t. $\boldsymbol{Z}$

$$E\left(\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{z}_n) + \log p(\boldsymbol{z}_n)\right)$$

$$= -\sum_{n=1}^{N}\{\frac{D}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2}E(\boldsymbol{z}_n)^T\boldsymbol{W}^T(\boldsymbol{x}_n - \boldsymbol{\mu})$$

$$+ \frac{1}{2\sigma^2}Tr(E(\boldsymbol{z}_n\boldsymbol{z}_n^T)\boldsymbol{W}^T\boldsymbol{W}) + \frac{1}{2}Tr(E(\boldsymbol{z}_n\boldsymbol{z}_n^T)) + \frac{M}{2}\log(2\pi)\}$$

- Noting that $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}_{\text{old}}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{M}_{\text{old}}^{-1}\boldsymbol{W}_{\text{old}}^T(\boldsymbol{x}-\bar{\boldsymbol{x}}), \sigma_{\text{old}}^2\boldsymbol{M}_{\text{old}}^{-1})$, we can readily evaluate

$$E(\boldsymbol{z}_n) = \boldsymbol{M}_{\text{old}}^{-1}\boldsymbol{W}_{\text{old}}^T(\boldsymbol{x}-\bar{\boldsymbol{x}})$$
$$E(\boldsymbol{z}_n\boldsymbol{z}_n^T) = \sigma_{\text{old}}^2\boldsymbol{M}_{\text{old}}^{-1} + E(\boldsymbol{z}_n)E(\boldsymbol{z}_n)^T$$

- In the M step, we find new estimates of $\boldsymbol{W}, \sigma^2$ that maximize the log-likelihood by setting their gradients to zero

$$\sigma_{\text{new}}^2 = \frac{1}{ND}\sum_{n=1}^{N}\{\|\boldsymbol{x}_n-\bar{\boldsymbol{x}}\|^2 - 2E(\boldsymbol{z}_n)^T\boldsymbol{W}_{\text{new}}^T(\boldsymbol{x}_n-\bar{\boldsymbol{x}})$$
$$+ Tr(E(\boldsymbol{z}_n\boldsymbol{z}_n^T)\boldsymbol{W}_{\text{new}}^T\boldsymbol{W}_{\text{new}})\}$$
$$\boldsymbol{W}_{\text{new}} = \left[\sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})E(\boldsymbol{z}_n)^T\right]\left[\sum_{n=1}^{N}E(\boldsymbol{z}_n\boldsymbol{z}_n^T)\right]^{-1}$$

- In computing the gradient w.r.t. a matrix $\boldsymbol{A}$, we make use of the following equality

$$\frac{\partial Tr(\boldsymbol{A}^T\boldsymbol{B})}{\partial \boldsymbol{A}} = \boldsymbol{B}$$

- The EM algorithm can be implemented in an on-line form, in which each data point is read in, processed, and then discarded before the next data point is considered

- The probabilistic PCA, together with the EM, allows us to handle missing data; the unobserved elements $\boldsymbol{x}_n^{(u)}$ of $\boldsymbol{x}_n$ can be marginalized in computing the corresponding likelihood

$$\int p(\boldsymbol{x}_n^{(o)}, \boldsymbol{x}_n^{(u)}, \boldsymbol{z}_n; \boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)d\boldsymbol{x}_n^{(u)} = p(\boldsymbol{x}_n^{(o)}, \boldsymbol{z}_n; \boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$$

Review

- Linear factor models: fully probabilistic models with latent variables
- Example: Probabilistic PCA
- Probabilistic PCA vs. standard PCA
- Learning low-dimensional manifolds
- Advantages of fully probabilistic models
- The EM algorithm for parameter estimation with latent variables