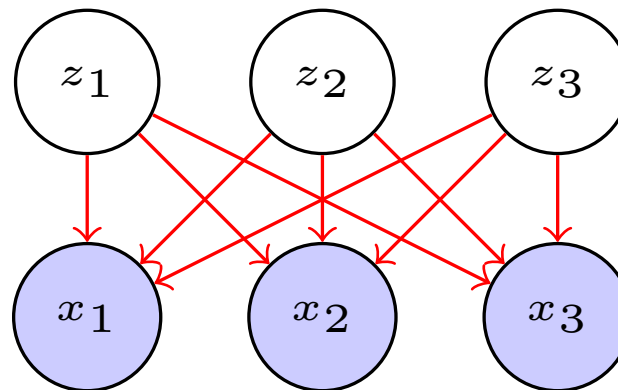


# Chapter 13

## Linear Factor Models

## Linear Factor Models

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



- In symbols, we have

$$\mathbf{z} \sim p(\mathbf{z})$$

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

- The latent variables  $\mathbf{z}$  capture the dependencies between the observed data  $\mathbf{x}$  and are known as explanatory factors
- Generally,  $p(\mathbf{z})$  is assumed to be factorial, i.e.,

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

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

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

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

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

- With these, we have a complete probabilistic model

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

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

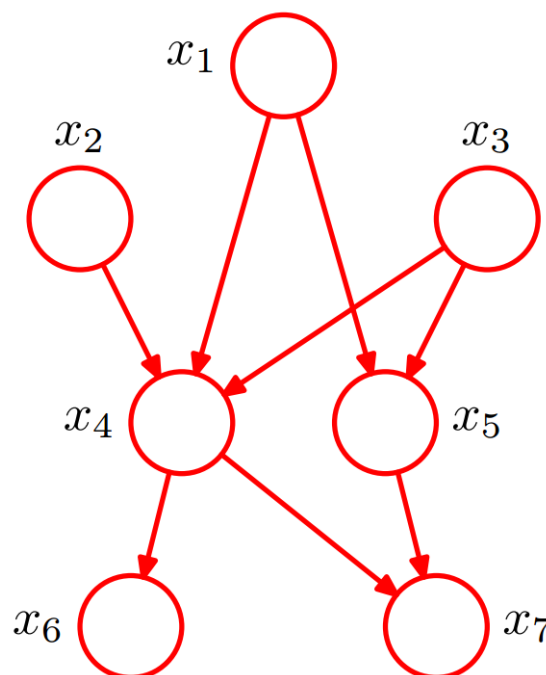
- In principle, we can
  - Do any probabilistic inference, e.g., to predict  $z$  based on  $x$

$$p(z|x) \propto p(z)p(x|z)$$

- Generate  $x$  by first sampling  $z$  and then using  $x = Wz + \mu + \epsilon$
- etc.

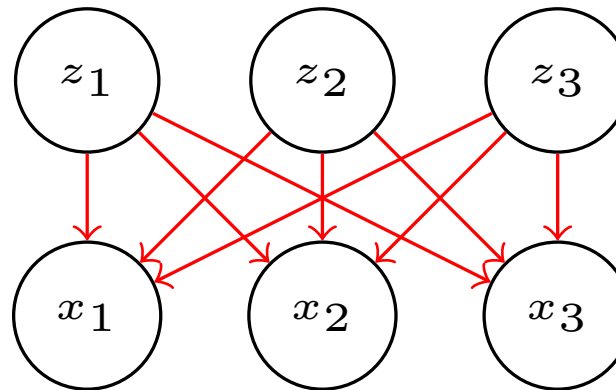
## Graphical Models 101

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



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

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



$$p(x_1, x_2, x_3, z_1, z_2, z_3) = p(z_1)p(z_2)p(z_3) \\ p(x_1|z_1, z_2, z_3)p(x_2|z_1, z_2, z_3)p(x_3|z_1, z_2, z_3)$$

- 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

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(x_1, x_2, x_3, z_1, z_2, z_3)}{p(z_1)p(z_2)p(z_3)} \\ = p(x_1|z_1, z_2, z_3)p(x_2|z_1, z_2, z_3)p(x_3|z_1, z_2, z_3)$$

## Probabilistic Principle Component Analysis (PCA)

---

- Probabilistic PCA is one example of linear factor models with

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$

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

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

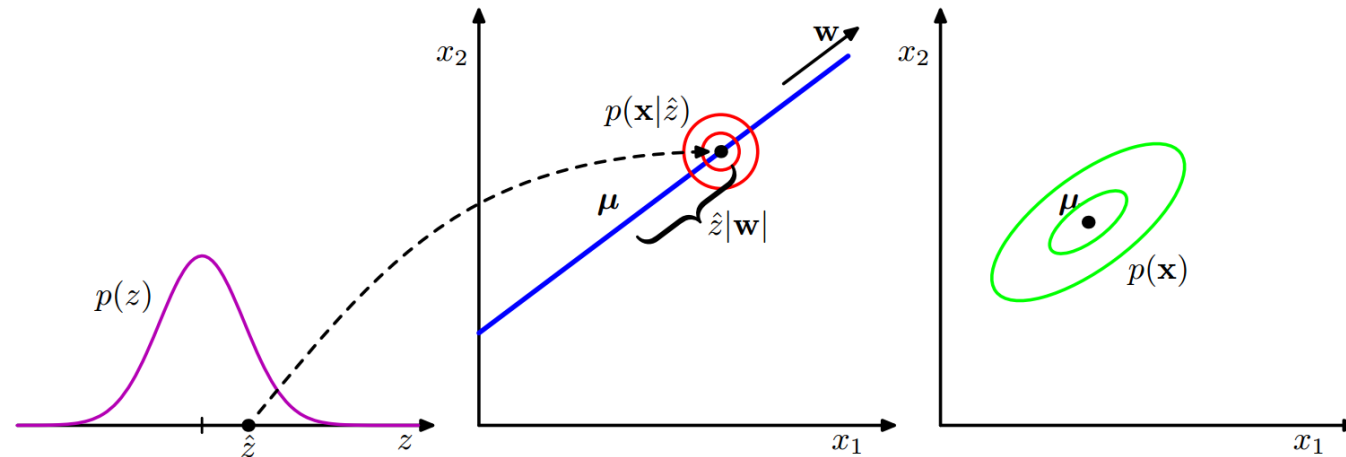
$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

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

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

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

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



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

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

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

$$\begin{aligned} E((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T) &= E((\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^T) \\ &= E(\mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T) + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \end{aligned}$$



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

- Remarks

- The resulting Gaussian distribution  $p(\mathbf{x})$  is governed by  $\boldsymbol{\mu}, \mathbf{W}, \sigma^2$ , 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  $\mathbf{R}\mathbf{R}^T = \mathbf{R}^T \mathbf{R} = \mathbf{I}$  to the latent space  $\tilde{\mathbf{z}} = \mathbf{R}\mathbf{z}$  does not change  $p(\mathbf{x})$ ; as can be seen,

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

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

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

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

where

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

- This is solved straightforwardly by observing that

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

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

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

## Maximum Likelihood PCA

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

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

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

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

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

$$\mathbf{W}_{\text{ML}} = \mathbf{U}(\mathbf{L} - \sigma_{\text{ML}}^2 \mathbf{I})^{1/2} \mathbf{R}$$

$$\sigma_{\text{ML}}^2 = \frac{1}{D - M} \sum_{i=M+1}^D \lambda_i$$

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

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

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

convenience)

- To summarize, we have a data model

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$

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

which gives

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

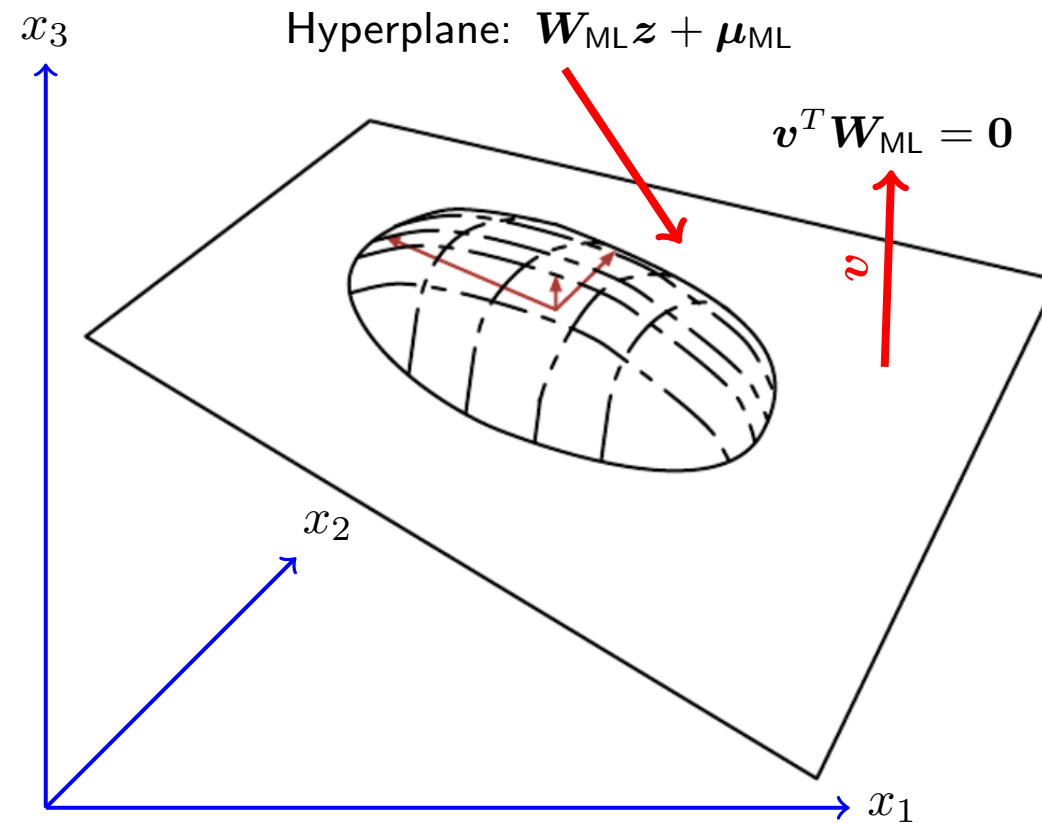
- Observations
  - Along the principle axes  $\mathbf{v} = \mathbf{U}_{:,i}$ , the model correctly captures the data variance

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

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

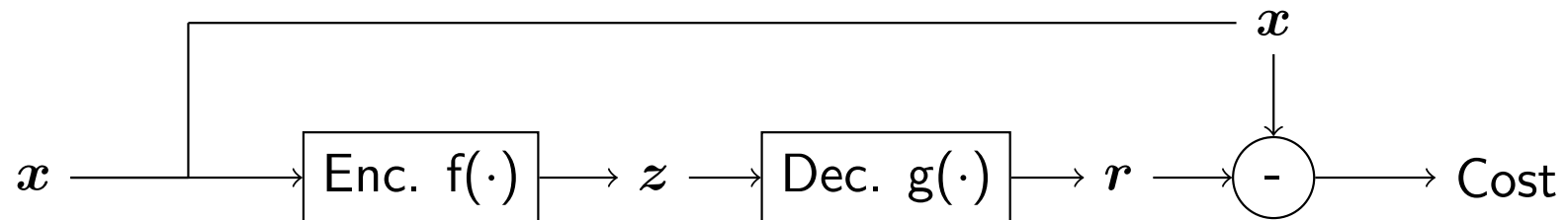
discarded eigenvalues

$$E[(\mathbf{v}^T(\mathbf{x} - \mathbf{u}_{\text{ML}}))^2] = \mathbf{v}^T \mathbf{C}_{\text{ML}} \mathbf{v} = \sigma_{\text{ML}}^2$$



## Standard PCA

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



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

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

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

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

which gives

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

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

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

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



where

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

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  $\mathbf{W}$  has its columns composed of the eigenvectors of the (scaled) sample covariance matrix  $\tilde{\mathbf{X}}^{(\text{train})} \tilde{\mathbf{X}}^{(\text{train})T}$  that correspond to the largest  $M$  eigenvalues

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

## Standard PCA vs. Probabilistic PCA

- Standard PCA: Deterministic encoder/decoder

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

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

- Probabilistic PCA: Stochastic encoder/decoder

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

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

where

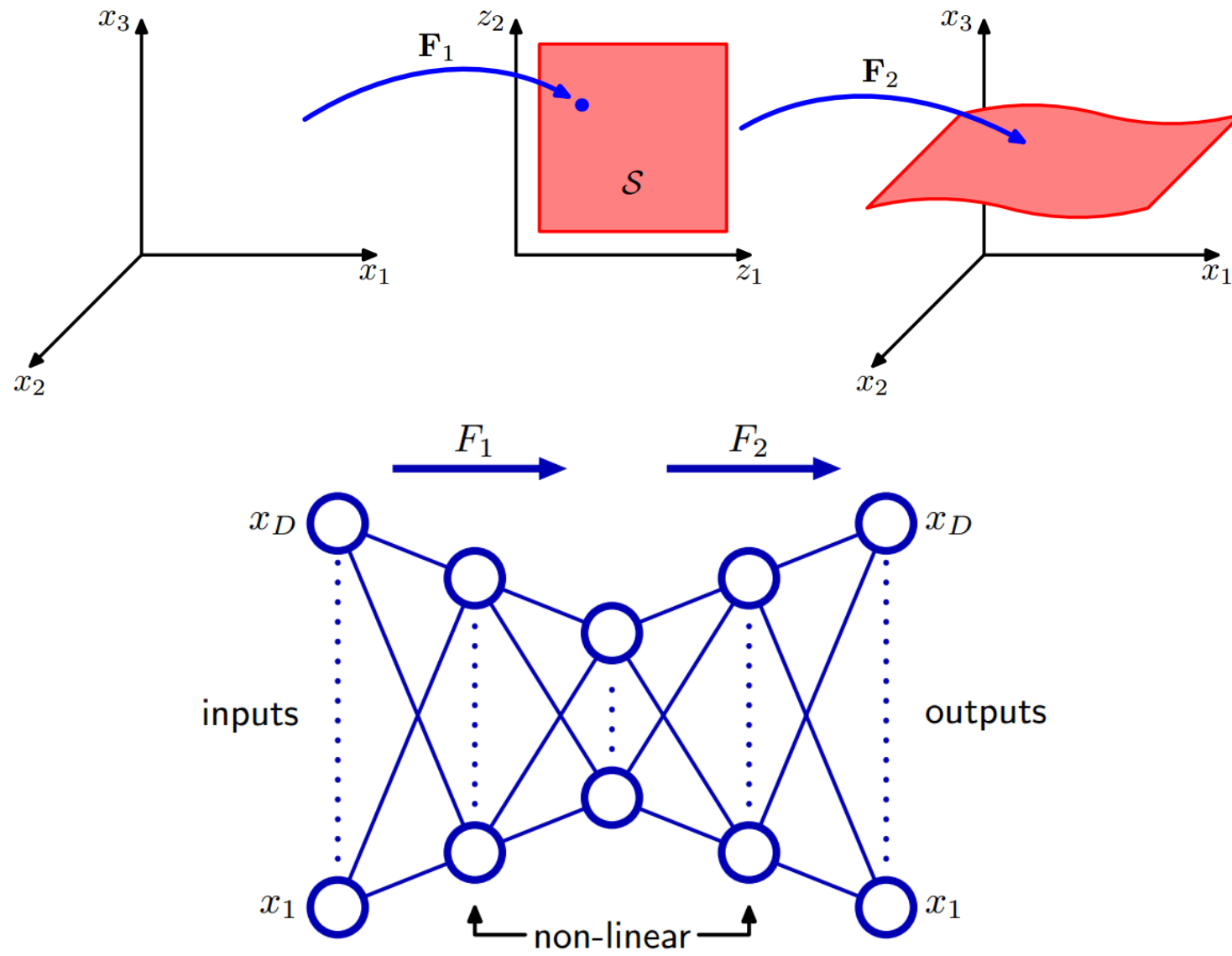
$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$$

- 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  $\mathbf{W}z + \bar{x}$
- The idea of dimension reduction can be extended to incorporate neural networks to learn a general, non-linear manifold



## The Expectation Maximization (EM) Algorithm

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} & \int q(\mathbf{Z}) \log p(\mathbf{X}; \boldsymbol{\theta}) d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\mathbf{Z} - \int q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) d\mathbf{Z} \\ &= \underbrace{\int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}}_{\text{ELBO}} \\ &+ \underbrace{\int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) d\mathbf{Z}}_{\text{KL Divergence}} \end{aligned}$$

to arrive at

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

where

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

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

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

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

with equality if and only if

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

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



- Now, by choosing deliberately

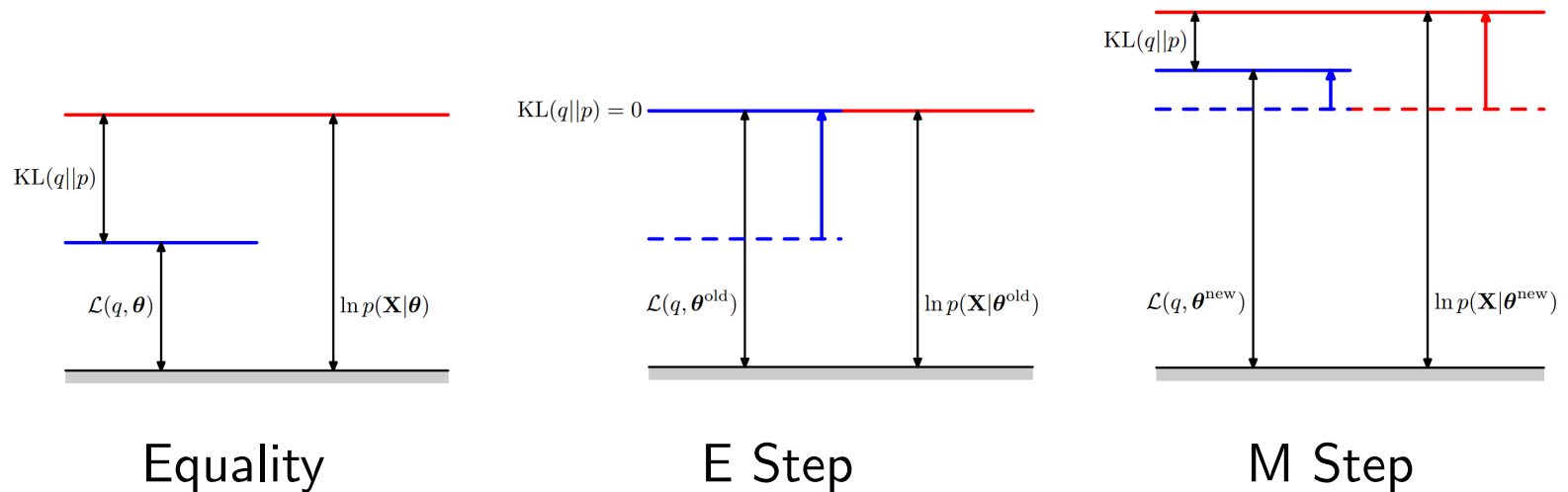
$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}^{\text{old}}),$$

we have

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

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

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



## EM for Probabilistic PCA

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

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

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

$$\begin{aligned} & E \left( \sum_{n=1}^N \log p(\mathbf{x}_n | \mathbf{z}_n) + \log p(\mathbf{z}_n) \right) \\ &= - \sum_{n=1}^N \left\{ \frac{D}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} E(\mathbf{z}_n)^T \mathbf{W}^T (\mathbf{x}_n - \boldsymbol{\mu}) \right. \\ &\quad \left. + \frac{1}{2\sigma^2} \text{Tr}(E(\mathbf{z}_n \mathbf{z}_n^T) \mathbf{W}^T \mathbf{W}) + \frac{1}{2} \text{Tr}(E(\mathbf{z}_n \mathbf{z}_n^T)) + \frac{M}{2} \log(2\pi) \right\} \end{aligned}$$

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

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

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

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

$$\begin{aligned} \sigma_{\text{new}}^2 = \frac{1}{ND} \sum_{n=1}^N \{ & \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 - 2E(\mathbf{z}_n)^T \mathbf{W}_{\text{new}}^T (\mathbf{x}_n - \bar{\mathbf{x}}) \\ & + \text{Tr}(E(\mathbf{z}_n \mathbf{z}_n^T) \mathbf{W}_{\text{new}}^T \mathbf{W}_{\text{new}}) \} \end{aligned}$$

$$\mathbf{W}_{\text{new}} = \left[ \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}}) E(\mathbf{z}_n)^T \right] \left[ \sum_{n=1}^N E(\mathbf{z}_n \mathbf{z}_n^T) \right]^{-1}$$

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

following equality

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

- 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  $\mathbf{x}_n^{(u)}$  of  $\mathbf{x}_n$  can be marginalized in computing the corresponding likelihood

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

## 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