Deep Generative Models
Probabilistic PCA

Fall Semester 2025

René Vidal

Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

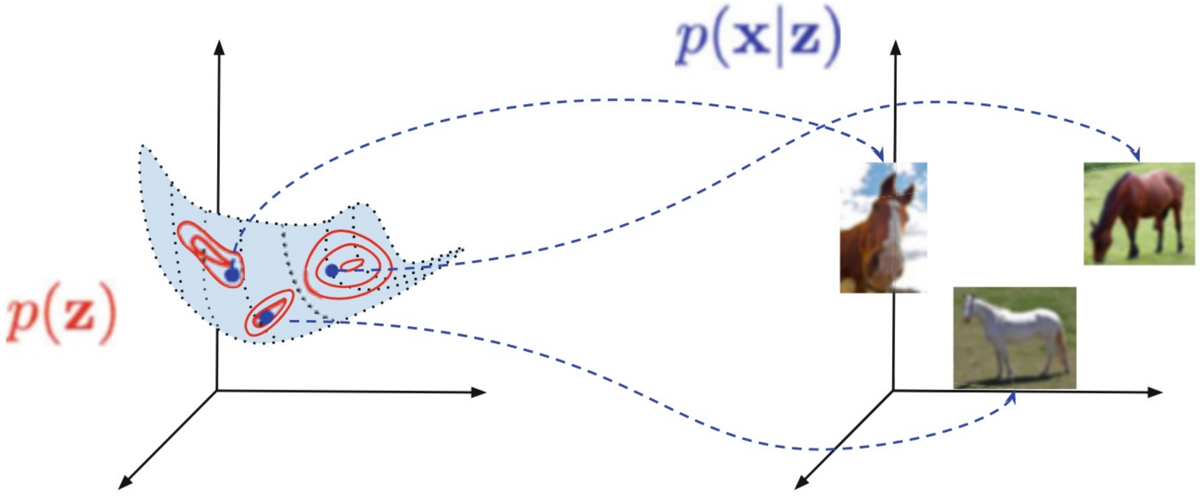
Penn
UNIVERSITY of PENNSYLVANIA

Outline

- Probabilistic PCA: Model
  - Joint, Marginal and Conditional Distributions
  - PPCA as an Encoder-Decoder Architecture
- Linear Algebra Background
  - Singular Value Decomposition (SVD)
  - SVD Properties
- Probabilistic PCA: Learning
  - Maximum Likelihood Estimation
- Applications
  - Application of PPCA to Generating Face Images

Latent Variable Models

- X = observed variable
- Z = latent variable



A latent variable model and a generative process. Note the low-dimensional manifold (here 2D) embedded in the high-dimensional space (here 3D)

- $\boldsymbol{z} \sim p(\boldsymbol{z})$
- $\boldsymbol{x} \sim p(\,\boldsymbol{x} \mid \boldsymbol{z}\,)$

- Factorization of the joint model

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

- Marginalization of the model

$$p(\boldsymbol{x}) = \int p(\,\boldsymbol{x} \mid \boldsymbol{z}\,)p(\boldsymbol{z})d\boldsymbol{z}$$

Probabilistic Principal Component Analysis: Model

- We consider continuous random variables only, i.e.,

$$\boldsymbol{z} \in \mathbb{R}^d \text{ and } \boldsymbol{x} \in \mathbb{R}^D \text{ with } d \ll D$$

- The distribution of $\boldsymbol{z}$ is the standard Gaussian, i.e.,

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}).$$

- The dependency between $\boldsymbol{z}$ and $\boldsymbol{x}$ is linear, plus Gaussian additive noise:

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\varepsilon}$$

- Here $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and independent from $\boldsymbol{z}$.

Probabilistic Principal Component Analysis: Model

- PPCA model

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \,.$$

- $\boldsymbol{x}$ is a linear combination of Gaussians, thus $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I})$ because

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}] &= \mathbb{E}[\boldsymbol{W}\boldsymbol{z}] + \mathbb{E}[\boldsymbol{b}] + \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{W}\mathbb{E}[\boldsymbol{z}] + \boldsymbol{b} + \boldsymbol{0} = \boldsymbol{b} \\
\mathbb{V}[\boldsymbol{x}] &= \mathbb{V}[\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}] = \boldsymbol{W}\mathbb{V}(\boldsymbol{z})\boldsymbol{W}^\top + \mathbb{V}[\boldsymbol{\epsilon}] = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}
\end{aligned}$$

- $\boldsymbol{x} \mid \boldsymbol{z}$ is a constant + a Gaussian, thus $p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \sigma^2 \boldsymbol{I})$ because

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{z}] &= \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} \\
\mathbb{V}[\boldsymbol{x} \mid \boldsymbol{z}] &= \mathbb{V}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}
\end{aligned}$$

Probabilistic Principal Component Analysis: Model

- PPCA model: $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon},\ p(\boldsymbol{z}) = \mathcal{N}(\ \boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}\ ),\ p(\boldsymbol{\epsilon}) = \mathcal{N}(\ \boldsymbol{\epsilon} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I}\ ),$

$$p(\ \boldsymbol{x} \mid \boldsymbol{z}\ ) = \mathcal{N}(\ \boldsymbol{x} \mid \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \sigma^2 \boldsymbol{I}\ ), \qquad p(\boldsymbol{x}) = \mathcal{N}(\ \boldsymbol{x} \mid \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}\ )$$

- Let $\boldsymbol{M} = \boldsymbol{W}^\top \boldsymbol{W} + \sigma^2 \boldsymbol{I}$. We can compute the conditional distribution of $(\ \boldsymbol{z} \mid \boldsymbol{x}\ )$ as

$$p(\ \boldsymbol{z} \mid \boldsymbol{x}\ ) = \frac{p(\ \boldsymbol{x} \mid \boldsymbol{z}\ ) p(\boldsymbol{z})}{p(\boldsymbol{x})} \propto e^{-\frac{1}{2\sigma^2}||\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{b}||^2} e^{-\frac{1}{2}||\boldsymbol{z}||^2}$$

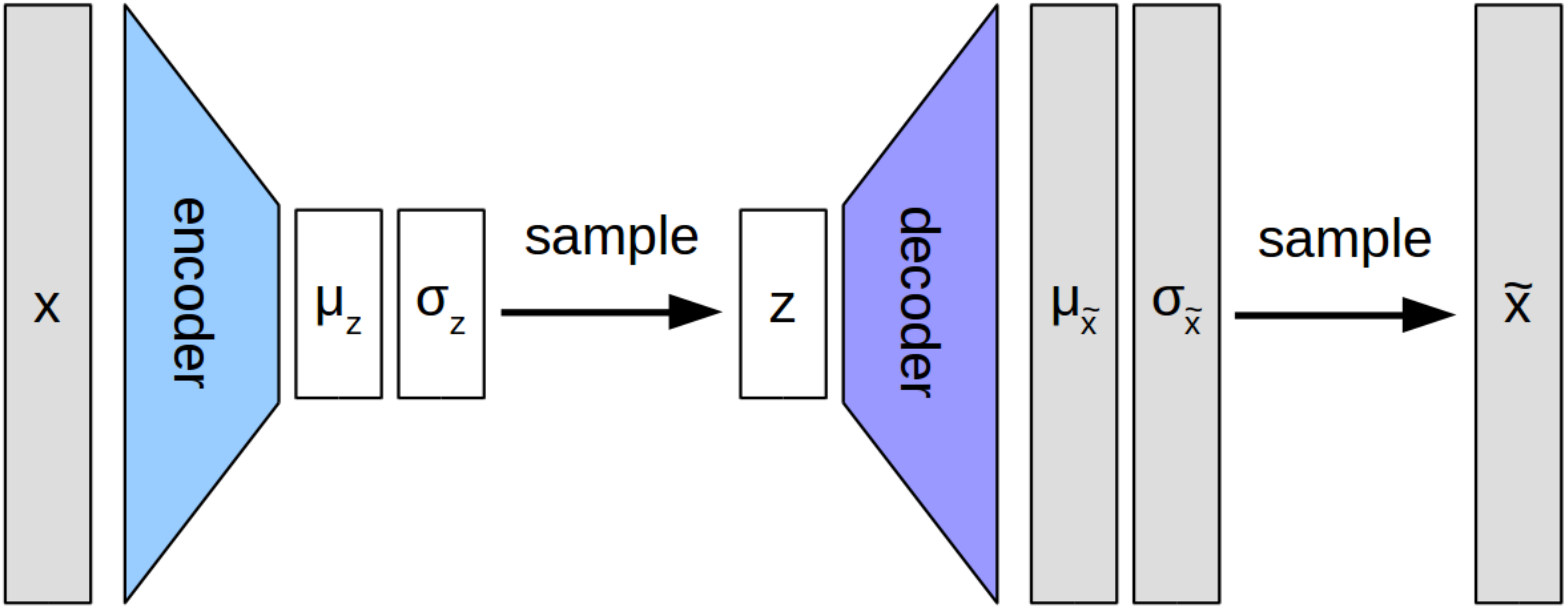
$$p(\ \boldsymbol{z} \mid \boldsymbol{x}\ ) \propto e^{-\frac{1}{2\sigma^2}\left(\boldsymbol{z}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{z} - 2\boldsymbol{z}^T \boldsymbol{W}^T (\boldsymbol{x} - \boldsymbol{b}) + \sigma^2 ||\boldsymbol{z}||^2\right)} \propto e^{-\frac{1}{2\sigma^2}\left(\boldsymbol{z}^T \boldsymbol{M} \boldsymbol{z} - 2\boldsymbol{z}^T \boldsymbol{W}^T (\boldsymbol{x} - \boldsymbol{b})\right)}$$

$$p(\ \boldsymbol{z} \mid \boldsymbol{x}\ ) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{M}^{-1} \boldsymbol{W}^\top (\boldsymbol{x} - \boldsymbol{b}), \sigma^2 \boldsymbol{M}^{-1})$$

PPCA as an Encoder Decoder Architecture

- PPCA model: $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon},\ p(\boldsymbol{z}) = \mathcal{N}(\ \boldsymbol{z} \mid \mathbf{0}, \boldsymbol{I}\ ),\ p(\boldsymbol{\epsilon}) = \mathcal{N}(\ \boldsymbol{\epsilon} \mid \mathbf{0}, \sigma^2\boldsymbol{I}\ ),$

$p(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z} \mid M^{-1}W^\top(\boldsymbol{x} - \boldsymbol{b}), \sigma^{-2}M)$ $\qquad$ $p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x} \mid W\boldsymbol{z} + \boldsymbol{b}, \sigma^2 I)$



Singular Value Decomposition (SVD)

- A matrix $X \in \mathbb{R}^{m\times n}$ of rank r can be decomposed as

$$X = U\Sigma V^\top = [u_1 \cdots u_r u_{r+1} \cdots u_m] \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} [v_1 \cdots v_r v_{r+1} \cdots v_m]^\top$$

- $\Sigma \in \mathbb{R}^{m\times n}$ is a “diagonal” matrix with r singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$
- $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthonormal matrices of left/right singular vectors
- Orthonormal means that columns are unit norm and orthogonal to each other:

$$U^\top U = UU^\top = I_{m\times m}, \qquad V^\top V = VV^\top = I_{n\times n}$$

Compact Singular Value Decomposition (SVD)

- A matrix $X \in \mathbb{R}^{m\times n}$ of rank r can be decomposed as

$$X = U\Sigma V^\top = [u_1 \cdots u_r u_{r+1} \cdots u_m] \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} [v_1 \cdots v_r v_{r+1} \cdots v_m]^\top$$

$$= U_1 \Sigma_1 V_1^\top = [u_1 \cdots u_r] \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} [v_1 \cdots v_r]^\top$$

- $U_1 \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ are orthonormal matrices of left/right singular vectors
- $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with r singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$

SVD Properties

- We can use any partition of the SVD of a matrix to expand it as

$$X = U\Sigma V^\top = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^\top = U_1 \Sigma_1 V_1^\top + U_2 \Sigma_2 V_2^\top$$

- **Example**: $I = U_1 U_1^\top + U_2 U_2^\top$, where $U_1^\top U_1 = I$, $U_1^\top U_2 = 0$, and $U_2^\top U_2 = I$

- If X is invertible, we can use its SVD to compute the inverse of a matrix

$$X = U\Sigma V^\top \Rightarrow X^{-1} = V\Sigma^{-1} U^\top$$

- We can use any partition of the SVD of a matrix to compute its inverse as

$$X = U_1 \Sigma_1 V_1^\top + U_2 \Sigma_2 V_2^\top \Rightarrow X^{-1} = V_1 \Sigma_1^{-1} U_1^\top + V_2 \Sigma_2^{-1} U_2^\top$$

SVD Properties

- We can use the SVD of a square matrix to compute its trace and determinant

$$X = U\Sigma V^\top \Rightarrow \mathrm{trace}(X) = \mathrm{trace}(\Sigma) = \sum_i \sigma_i \quad \text{and} \quad \det(X) = \det(\Sigma) = \prod_i \sigma_i$$

- We can use any partition of the SVD to compute its trace and determinant

$$X = U_1\Sigma_1 V_1^\top + U_2\Sigma_2 V_2^\top \Rightarrow \mathrm{trace}(X) = \mathrm{trace}(\Sigma_1) + \mathrm{trace}(\Sigma_2)$$

$$X = U_1\Sigma_1 V_1^\top + U_2\Sigma_2 V_2^\top \Rightarrow \det(X) = \det(\Sigma_1)\det(\Sigma_2)$$

- Example

$$X = \begin{bmatrix} I_m & 0 \\ 0 & \sigma^2 I_n \end{bmatrix} \Rightarrow \mathrm{trace}(X) = m + \sigma^2 n \quad \text{and} \quad \det(X) = \sigma^{2n}$$

Probabilistic Principal Component Analysis: Learning

- Recall the ML estimators of the parameters of a Gaussian $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are

$$\boldsymbol{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\,, \qquad \boldsymbol{\Sigma}_N = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}_N)(\boldsymbol{x}_i - \boldsymbol{\mu}_N)^T$$

- For PPCA we need to estimate the parameters of a Gaussian with structured covariance $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \sigma^2\boldsymbol{I}$. The estimate of the mean is the same as before $\boldsymbol{\mu} = \boldsymbol{\mu}_N$. To estimate $\boldsymbol{W}$, we need to maximize the log-likelihood w.r.t. $(\boldsymbol{W}, \sigma)$

$$\ell = -\frac{N}{2}\log(\det(\boldsymbol{\Sigma})) \; - \frac{N}{2}\mathrm{trace}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_N)$$

- Taking derivatives w.r.t. $\boldsymbol{W}$ we get

$$\frac{\partial \ell}{\partial \boldsymbol{W}} = \frac{\partial \ell}{\partial \boldsymbol{\Sigma}}\frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{W}} = -\frac{N}{2}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_N\boldsymbol{\Sigma}^{-1})2\boldsymbol{W} = 0 \;\Rightarrow\; \boldsymbol{\Sigma}_N\boldsymbol{\Sigma}^{-1}\boldsymbol{W} = \boldsymbol{W}$$

Probabilistic Principal Component Analysis: Learning

- We thus need to solve the nonlinear equations

$$\boldsymbol{\Sigma}_N \boldsymbol{\Sigma}^{-1} \boldsymbol{W} = \boldsymbol{W} \text{ and } \boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}$$

- A trivial solution is $(\boldsymbol{W}, \sigma) = (\boldsymbol{0}, 0)$, but this is a minimum of the log-likelihood.
- Another solution is $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_N$, but this would require the structure of the sample covariance $\boldsymbol{\Sigma}_N$ to match the structure of $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}$, i.e., the smallest eigenvalues would need to be all equal to each other and equal to σ^2.
- Alternatively, let

$$\boldsymbol{W} = [\boldsymbol{Z}_1 \quad \boldsymbol{Z}_2] \begin{bmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} [\boldsymbol{V}_1 \quad \boldsymbol{V}_2]^T = \boldsymbol{Z}_1 \boldsymbol{\Gamma}_1 \boldsymbol{V}_1^T$$

- Then

$$\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I} = \boldsymbol{Z}_1 \boldsymbol{\Gamma}_1^2 \boldsymbol{Z}_1^T + \sigma^2 (\boldsymbol{Z}_1 \boldsymbol{Z}_1^T + \boldsymbol{Z}_2 \boldsymbol{Z}_2^T) = \boldsymbol{Z}_1 (\boldsymbol{\Gamma}_1^2 + \sigma^2 \, \boldsymbol{I}) \boldsymbol{Z}_1^T + \sigma^2 \boldsymbol{Z}_2 \boldsymbol{Z}_2^T$$

$$\boldsymbol{\Sigma}^{-1} \boldsymbol{W} = (\boldsymbol{Z}_1 (\boldsymbol{\Gamma}_1^2 + \sigma^2 \, \boldsymbol{I})^{-1} \boldsymbol{Z}_1^T + \sigma^{-2} \boldsymbol{Z}_2 \boldsymbol{Z}_2^T) \boldsymbol{Z}_1 \boldsymbol{\Gamma}_1 \boldsymbol{V}_1^T = \boldsymbol{Z}_1 (\boldsymbol{\Gamma}_1^2 + \sigma^2 \, \boldsymbol{I})^{-1} \boldsymbol{\Gamma}_1 \boldsymbol{V}_1^T$$

Probabilistic Principal Component Analysis: Learning

- Therefore,

$$\boldsymbol{\Sigma}_N \boldsymbol{\Sigma}^{-1} \boldsymbol{W} = \boldsymbol{W} \Rightarrow \boldsymbol{\Sigma}_N \boldsymbol{Z}_1 (\boldsymbol{\Gamma}_1^2 + \sigma^2\, \boldsymbol{I})^{-1} \boldsymbol{\Gamma}_1 \boldsymbol{V}_1^T = \boldsymbol{Z}_1 \boldsymbol{\Gamma}_1 \boldsymbol{V}_{\mathbf{1}}^T \Rightarrow$$

$$\boldsymbol{\Sigma}_N \boldsymbol{Z}_1 (\boldsymbol{\Gamma}_1^2 + \sigma^2 \boldsymbol{I})^{-1} = \boldsymbol{Z}_1 \;\Rightarrow\; \boldsymbol{\Sigma}_N \boldsymbol{Z}_1 = \boldsymbol{Z}_1\, (\boldsymbol{\Gamma}_1^2 + \sigma^2 \boldsymbol{I}) \;\Rightarrow\; \boldsymbol{\Sigma}_N \boldsymbol{z}_i = \left(\gamma_i^2 + \sigma^2\right) \boldsymbol{z}_i$$

  - In other words, $\boldsymbol{z}_i$ is an eigenvector of $\boldsymbol{\Sigma}_N$ with eigenvalue $\gamma_i^2 + \sigma^2$.

- Thus, if $\boldsymbol{\Sigma}_N = [\boldsymbol{U}_1 \;\; \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{bmatrix} [\boldsymbol{U}_1 \;\; \boldsymbol{U}_2]^T$, then $\boldsymbol{Z}_1 = \boldsymbol{U}_1, \boldsymbol{\Gamma}_1^2 + \sigma^2 \boldsymbol{I} = \boldsymbol{\Lambda}_1$
  - In other words, $\boldsymbol{U}_1$ is a matrix whose d columns correspond to d singular vectors of $\boldsymbol{\Sigma}_N$

- Therefore, $\boldsymbol{W} = \boldsymbol{Z}_1 \boldsymbol{\Gamma}_1 \boldsymbol{V}_1^T = \boldsymbol{U}_1 (\boldsymbol{\Lambda}_1 - \sigma^2 \boldsymbol{I})^{\mathbf{1/2}} \boldsymbol{V}_1^T$

- Having “almost” found $\boldsymbol{W}$ (we don’t know which d columns), we now turn to finding σ.

Probabilistic Principal Component Analysis: Learning

- Recall the log-likelihood

$$\ell = -\frac{N}{2}\log(\det(\boldsymbol{\Sigma})) \ -\frac{N}{2}\mathrm{trace}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_N)$$

- We have $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \sigma^2 I$ and $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{\Gamma}_1\boldsymbol{V}_1^T$. Thus, $\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} = \boldsymbol{U}_1\boldsymbol{\Gamma}_1^{\boldsymbol{2}}\boldsymbol{U}_{\boldsymbol{1}}^{\boldsymbol{T}}$ and

$$\boldsymbol{\Sigma} \ = \boldsymbol{U}_1(\boldsymbol{\Gamma}_1^2 + \sigma^2\ \boldsymbol{I})\boldsymbol{U}_1^T + \sigma^2\boldsymbol{U}_2\boldsymbol{U}_2^T = \boldsymbol{U}_1\boldsymbol{\Lambda}_{\boldsymbol{1}}\boldsymbol{U}_1^T + \sigma^2\boldsymbol{U}_2\boldsymbol{U}_2^T$$

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_N = \left(\boldsymbol{U}_1\boldsymbol{\Lambda}_{\boldsymbol{1}}^{-\boldsymbol{1}}\boldsymbol{U}_1^T + \sigma^{-2}\boldsymbol{U}_2\boldsymbol{U}_2^T\right)(\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^T + \boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2^T) = \boldsymbol{U}_1\boldsymbol{U}_1^{\boldsymbol{T}} + \sigma^{-2}\boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2^T$$

- Substituting into the log-likelihood, we get

$$\ell = -\frac{N}{2}\log(\det(\boldsymbol{\Lambda}_1)\sigma^{2(D-d)}) \ -\frac{N}{2}(d + \sigma^{-2}\mathrm{trace}(\boldsymbol{\Lambda}_2))$$

Probabilistic Principal Component Analysis: Learning

- Taking the derivative yields $\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2}\left(\frac{D-d}{\sigma^2} - \frac{\mathrm{trace}(\Lambda_2)}{\sigma^4}\right) = 0 \Longrightarrow \sigma^2 = \frac{\mathrm{trace}(\Lambda_2)}{D-d}$
- Final piece: we have not shown that $\boldsymbol{\Lambda}_1$ corresponds to **top** d eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{N}}$
  - Exercise 2.13 in GPCA textbook

- **Theorem.** The ML estimates for the parameters of the PPCA model $\boldsymbol{b}, \boldsymbol{W}$, and σ can be obtained from the ML estimates of the mean and covariance of the data, $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_{\mathrm{N}}$, respectively, as

$$\boldsymbol{b} = \boldsymbol{\mu}_N,\ \boldsymbol{W} = \boldsymbol{U}_1(\boldsymbol{\Lambda}_1 - \sigma^2 I)^{1/2}\boldsymbol{R} \ \text{ and } \ \sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D} \lambda_i$$

- where $\boldsymbol{U}_1$ is the matrix with the top d eigenvectors of $\boldsymbol{\Sigma}_{\mathrm{N}}$, $\boldsymbol{\Lambda}_1$ is the matrix with the corresponding top d eigenvalues, $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ is an arbitrary orthogonal matrix, and λ_i is the ith largest eigenvalue of $\boldsymbol{\Sigma}_{\mathrm{N}}$.

Application of PPCA to Generating Face Images

Fig. 2.2 Face images of subject 20 under 10 different illumination conditions in the extended Yale B data set. All images are frontal faces cropped to size 192×168.

Application of PPCA to Generating Face Images

(a) mean face (b) first eigenface (c) second eigenface

Fig. 2.5 Mean face and the first two eigenfaces by applying PPCA to the ten images in Figure 2.2.

Application of PPCA to Generating Face Images

(a) Variation along the first eigenface

(b) Variation along the second eigenface

Fig. 2.6 Variation of the face images along the two eigenfaces given by PPCA. Each row plots $\boldsymbol{\mu} + y_i \boldsymbol{u}_i$ for $y_i = -1 : \frac{1}{3} : 1$, $i = 1, 2$.