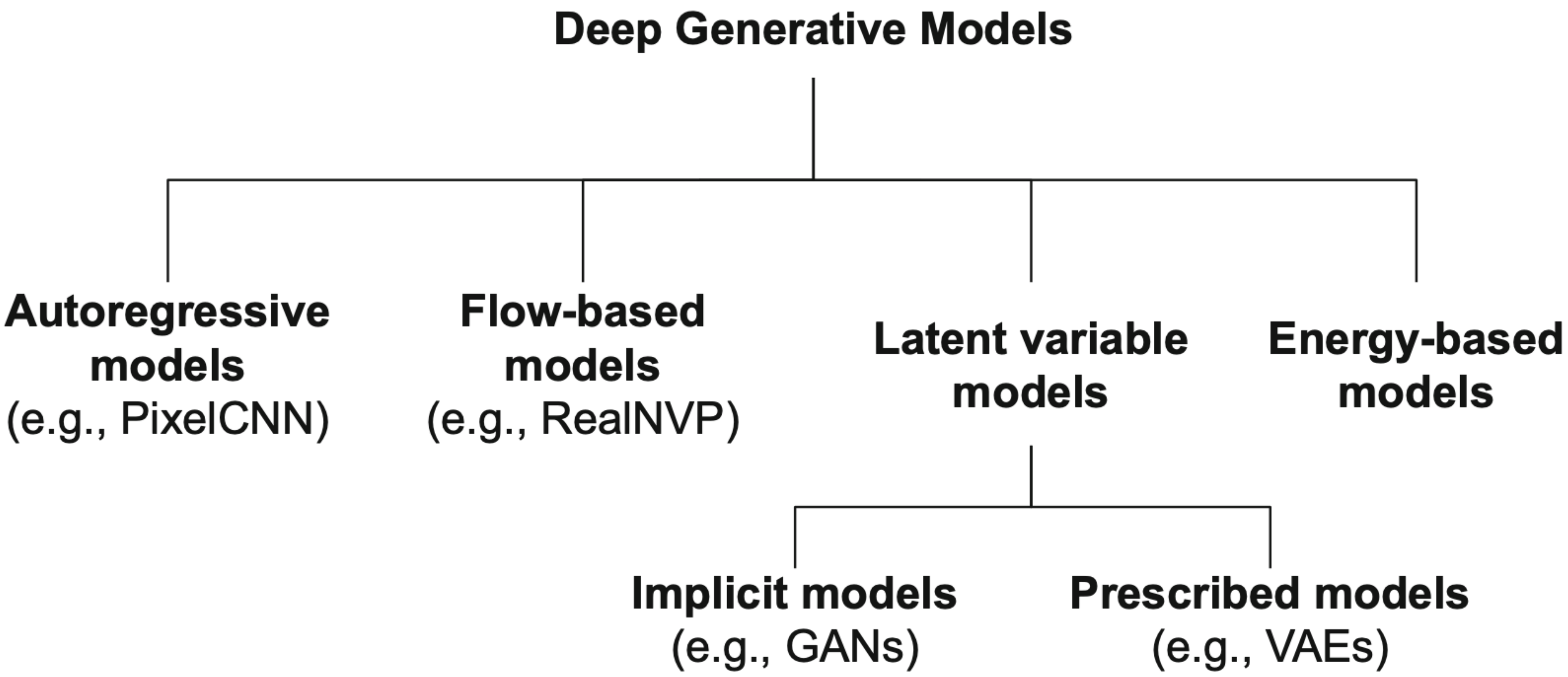
Deep Generative Models: Probabilistic PCA

Fall Semester 2024

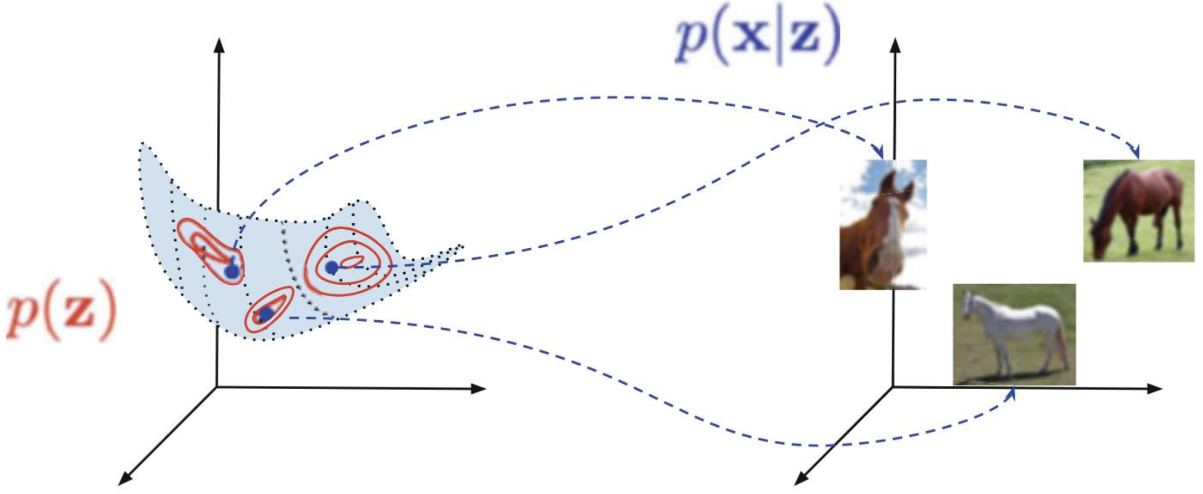
René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

Taxonomy of Generative Models



Latent Variable Models

- X = observed variable
- Z = latent variable



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

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

- Factorization of the joint model

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

- Marginalization of the model

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

Probabilistic Principal Component Analysis: Model

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

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

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

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

- The dependency between $\boldsymbol{z}$ and $\boldsymbol{x}$ is linear and we assume a Gaussian additive noise:

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

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

Probabilistic Principal Component Analysis: Model

- PPCA Model

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

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

$$E[\boldsymbol{x}] = E[\boldsymbol{W}\boldsymbol{z}] + \boldsymbol{b} + E[\boldsymbol{\epsilon}] = \boldsymbol{W}E[\boldsymbol{z}] + \boldsymbol{b} + \boldsymbol{0} = \boldsymbol{b}$$
$$Cov[\boldsymbol{x}] = Cov[\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}] = \boldsymbol{W}Cov(\boldsymbol{z})\boldsymbol{W}^\top + Cov[\boldsymbol{\epsilon}] = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}$$

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

$$E[\boldsymbol{x}|\boldsymbol{z}] = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + E[\boldsymbol{\epsilon}] = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$$
$$Cov[\boldsymbol{x}|\boldsymbol{z}] = Cov[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}$$

Probabilistic Principal Component Analysis: Model

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

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

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

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

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

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

Probabilistic Principal Component Analysis: Learning

- Recall the ML estimators of the parameters of a Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ 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 $\boldsymbol{W}$, recall the log-likelihood

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

- Taking derivatives w.r.t. 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 this 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 $\boldsymbol{W} = 0$, 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, $\boldsymbol{\Sigma}_N \boldsymbol{\Sigma}^{-1} \boldsymbol{W} = \boldsymbol{W} \Rightarrow \boldsymbol{\Sigma}_N \boldsymbol{Z}_1 (\boldsymbol{\Gamma}_1^2 + \sigma^2\ \boldsymbol{I})^{-1} \boldsymbol{\Gamma}_1 \boldsymbol{V}_1^T = \boldsymbol{Z}_1 \boldsymbol{\Gamma}_1 \boldsymbol{V}_{\mathbf{1}}^T$

- This leads to

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

- 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 $\boldsymbol{Z}_1 = \boldsymbol{U}_1, \boldsymbol{\Gamma}_1^2 + \sigma^2 \boldsymbol{I} = \boldsymbol{\Lambda}_1$.
- 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 now need to find σ.

Probabilistic Principal Component Analysis: Learning

- Recall the log-likelihood

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

- We have $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \sigma^2 I$, and $\boldsymbol{W} = \boldsymbol{Z}_1\boldsymbol{\Gamma}_1\boldsymbol{V}_1^T$ so

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

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

- Substituting into the log-likelihood, we get

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

Probabilistic Principal Component Analysis: Learning

- Taking the derivative yields $\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2}\left(\frac{D-d}{\sigma^2} - \frac{\mathrm{trace}(\Lambda_2)}{\sigma^4}\right) = 0 \Longrightarrow \sigma^2 = \frac{\mathrm{trace}(\Lambda_2)}{D-d}$

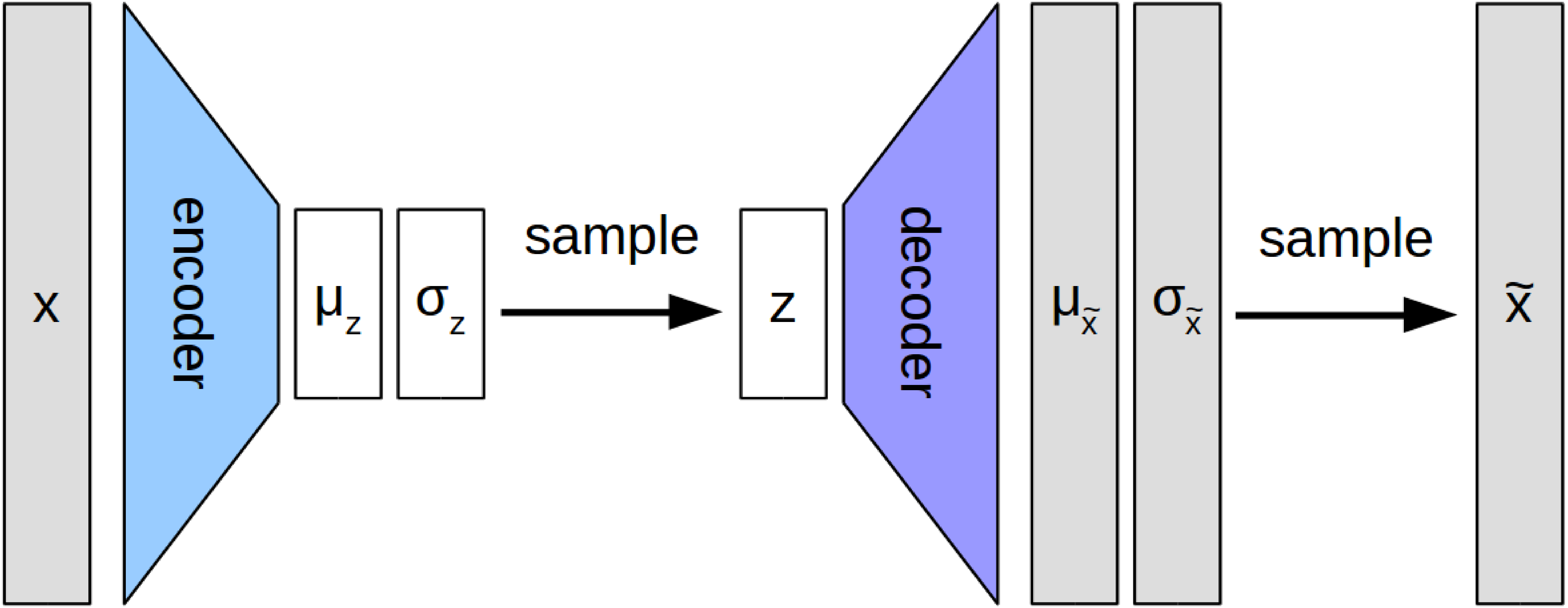
- **Theorem** The ML estimates for the parameters of the PPCA model μ, B, and σ can be obtained from the ML estimates of the mean and covariance of the data, $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_{\mathrm{N}}$, respectively, as

$$\boldsymbol{\mu} = \boldsymbol{\mu}_N,\ \boldsymbol{W} = U_1(\Lambda_1 - \hat{\sigma}^2 I)^{1/2} R \ \text{ and } \ \sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D} \lambda_i$$

- where U_1 is the matrix with the top d eigenvectors of $\boldsymbol{\Sigma}_{\mathrm{N}}$, Λ_1 is the matrix with the corresponding top d eigenvalues, $R \in \mathbb{R}^{d\times d}$ is an arbitrary orthogonal matrix, and λ_i is the ith largest eigenvalue of $\boldsymbol{\Sigma}_{\mathrm{N}}$.

PPCA as an Encoder Decoder Architecture

- $p(z|x) = N\left(M^{-1}W^{\top}(x-b), \sigma^{-2}M\right)$ $\qquad p(x|z) = N(x|Wz+b, \sigma^2 I)$



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 $y_i = -1 : \frac{1}{3} : 1$, $i = 1, 2$.