Redução de Dimensionalidade

Prof. Danilo Silva

EEL7514/EEL7513 - Tópico Avançado em Processamento de Sinais
EEL410250 - Aprendizado de Máquina

EEL / CTC / UFSC

Introdução

Redução de Dimensionalidade

- Consiste em aplicar uma transformação

$$f : \mathbf{x} \in \mathbb{R}^n \mapsto \mathbf{z} \in \mathbb{R}^K \quad \textit{(encoding)}$$

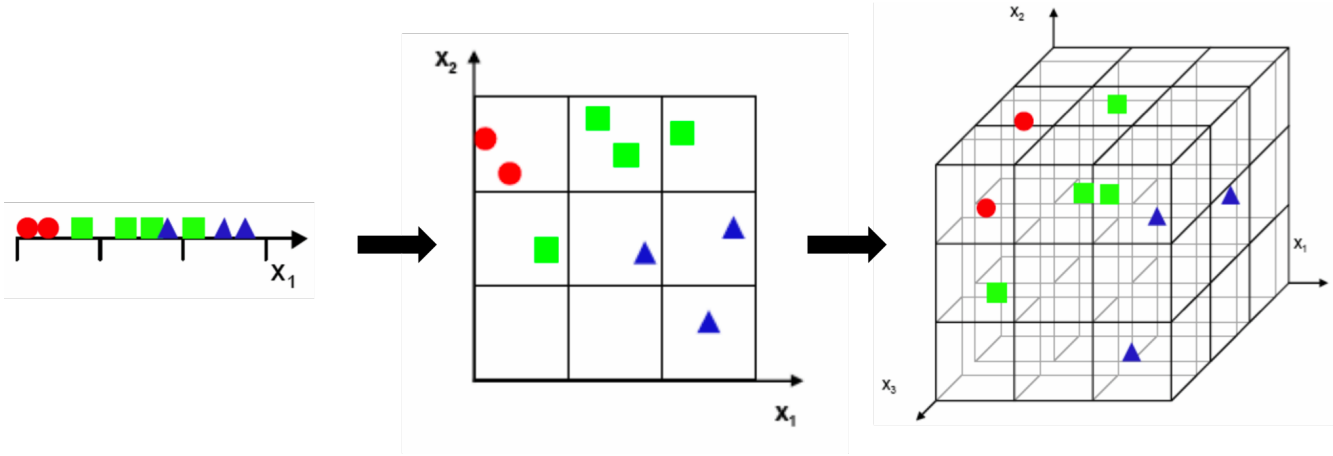
onde $K < n$, de tal forma a permitir uma reconstrução aproximada

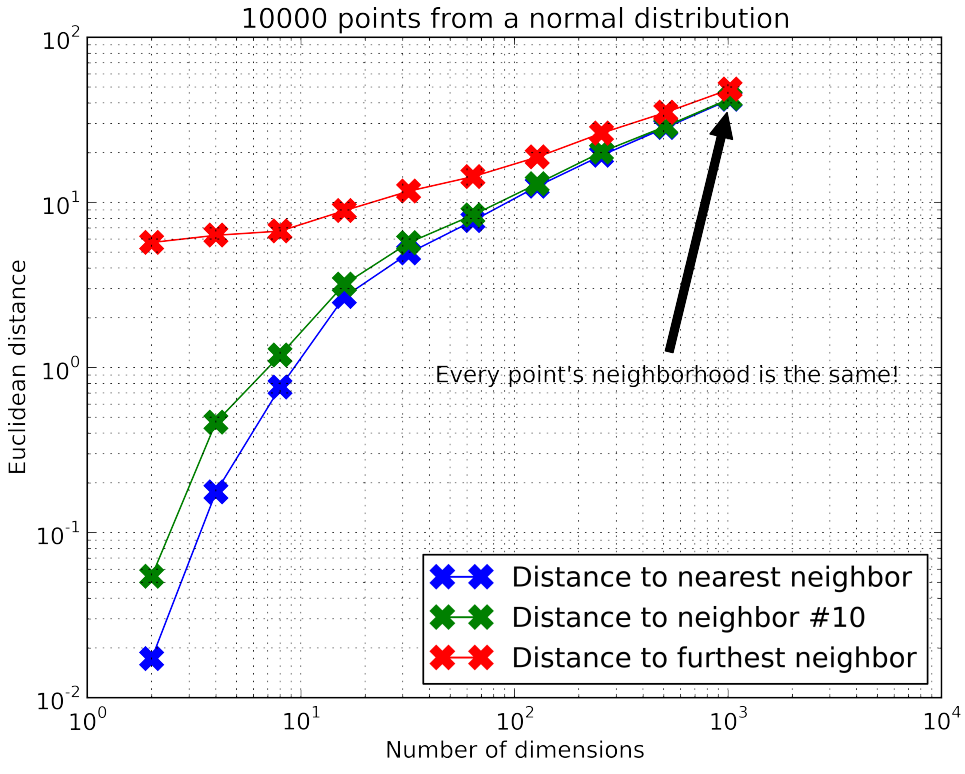
$$g : \mathbf{z} \in \mathbb{R}^K \mapsto \hat{\mathbf{x}} \in \mathbb{R}^n \quad \textit{(decoding)}$$

onde $\hat{\mathbf{x}} \approx \mathbf{x}$ sob algum critério (ex: $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \approx 0$).

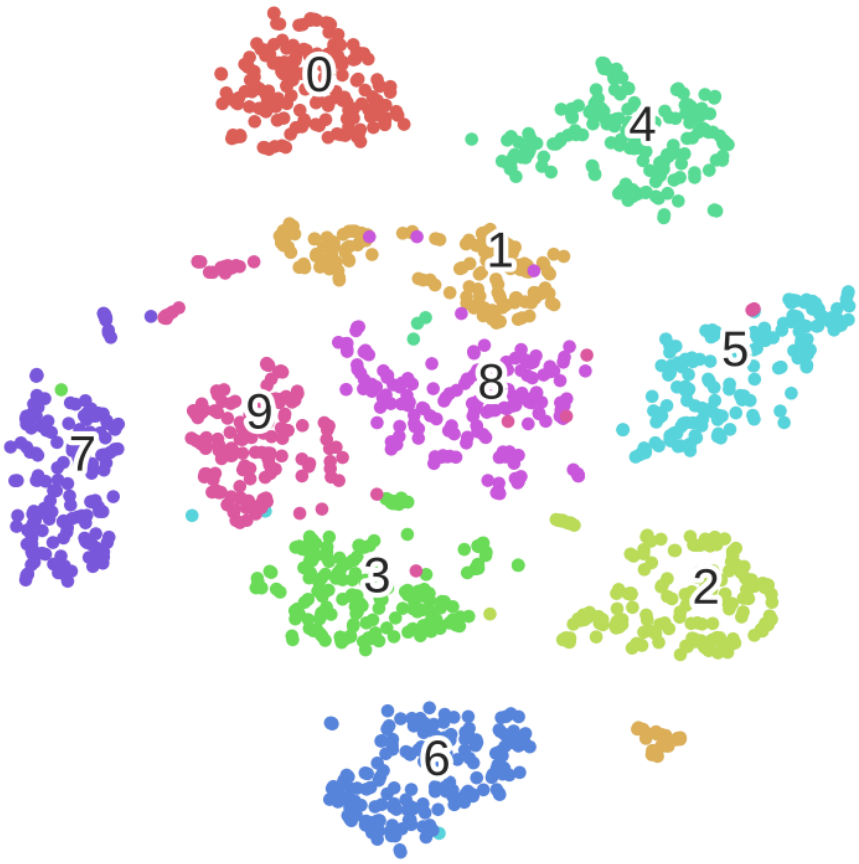
- Variáveis $z_1, \ldots, z_K$ são chamadas de latentes (não-observáveis)
- Motivação/aplicações:
  - Acelerar algoritmos de aprendizado (ex: clustering)
  - Amenizar problemas causados por dimensionalidade elevada (*curse of dimensionality*) (ex: concentração de distâncias, *overfitting*)
  - Detecção de anomalias (ex: $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 > \epsilon \implies$ anomalia)
  - Facilitar a visualização (geralmente em 2D)

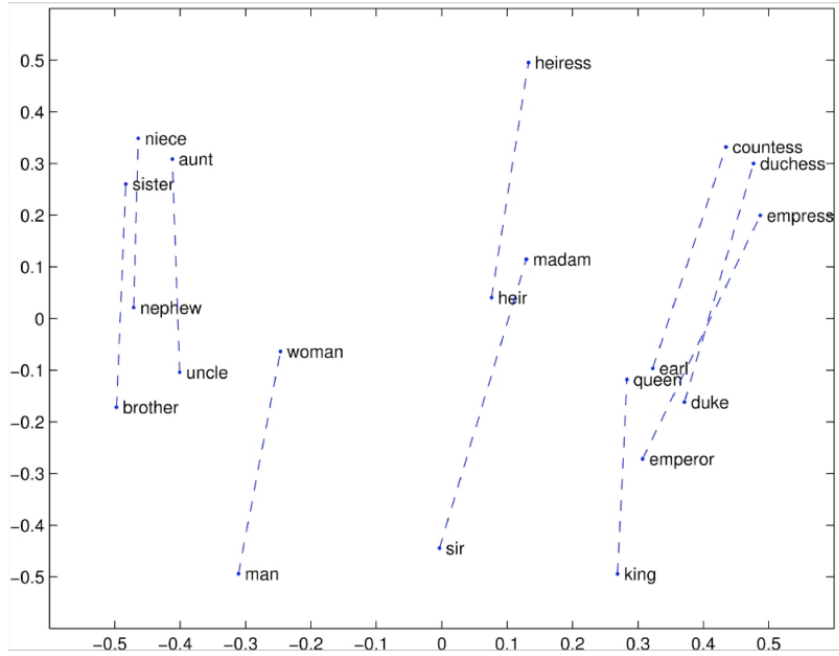
Exemplo: *Curse of Dimensionality*





Exemplo: Visualização em 2D





Exemplo: Dimensão Intrínseca

- Pode ser representado com apenas 2 variáveis latentes

Problema de Otimização

- Conjunto de dados (não-rotulados): $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$
- Dimensão do espaço latente: K
- Função custo:

$$J = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2, \quad \hat{\mathbf{x}}^{(i)} = g(\mathbf{z}^{(i)}), \quad \mathbf{z}^{(i)} = f(\mathbf{x}^{(i)})$$

  onde $f : \mathbb{R}^n \to \mathbb{R}^K$ e $g : \mathbb{R}^K \to \mathbb{R}^n$

- Problema de otimização (caso genérico):

$$\min_{f,g} \; J$$

Análise de Componentes Principais

Análise de Componentes Principais

Principal Component Analysis (PCA)

- Suponha que os atributos $x_1, \ldots, x_n$ estão centralizados (média nula)

$$\mu_{x_j} \triangleq \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} = 0$$

possivelmente após pré-processamento: $x_j^{(i)} \leftarrow x_j^{(i)} - \mu_{x_j}$

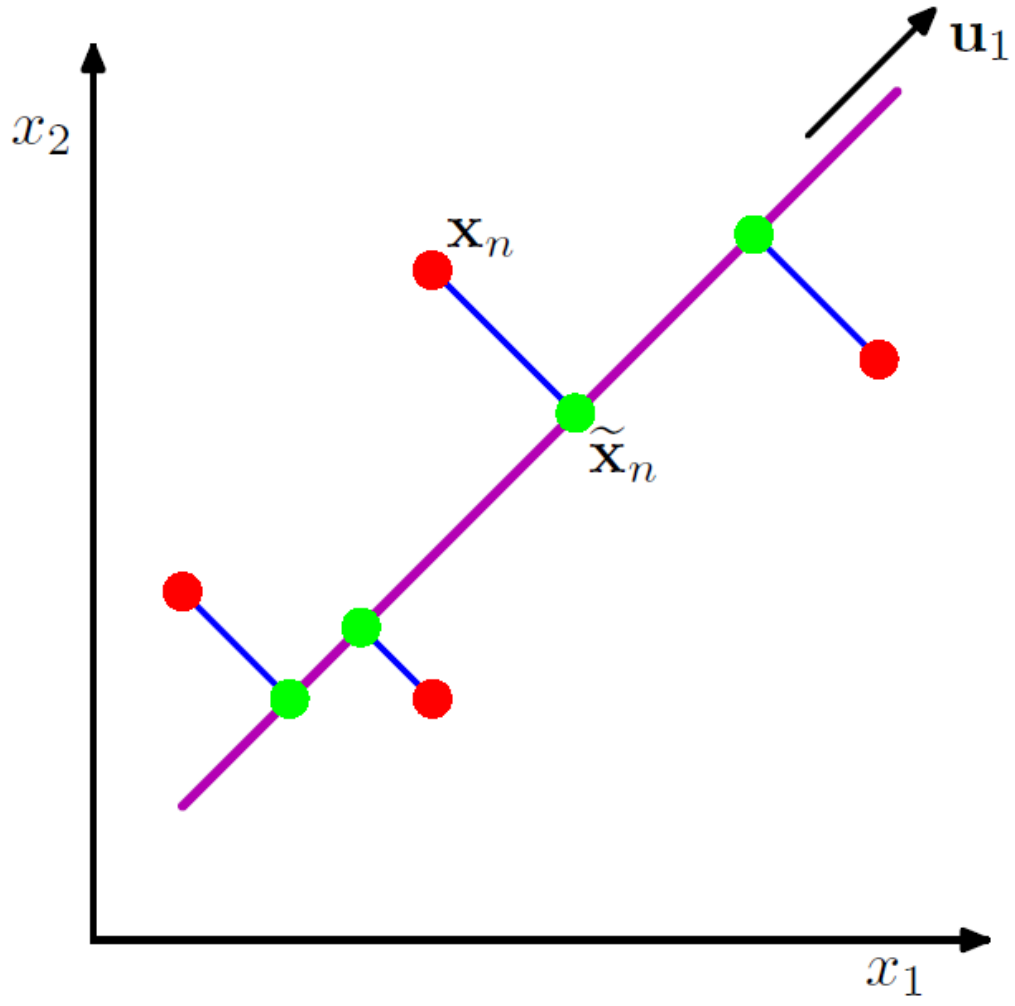
- PCA utiliza para decodificação uma transformação linear:

$$\hat{\mathbf{x}}^{(i)} = \mathbf{G}\mathbf{z}^{(i)} = \sum_{k=1}^{K} \mathbf{g}_k z_k^{(i)}, \quad \mathbf{G} = \begin{bmatrix} | & & | \\ \mathbf{g}_1 & \cdots & \mathbf{g}_K \\ | & & | \end{bmatrix}, \quad \mathbf{g}_k \in \mathbb{R}^n$$

- Função custo:

$$J = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2 = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \mathbf{G}\mathbf{z}^{(i)}\|^2$$

Interpretação Geométrica



- Reconstrução $\hat{\mathbf{x}}^{(i)}$ pertence ao subespaço gerado pelas colunas de $\mathbf{G}$

Formulação Matemática

- Solução ótima (para $\mathbf{G}$ fixo):

$$\mathbf{z}^{(i)} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}^{(i)}$$

- Como o custo só depende do subespaço, podemos, sem perda de generalidade, considerar $\mathbf{g}_1, \ldots, \mathbf{g}_K$ uma base ortonormal, isto é,

$$\mathbf{g}_i^T\mathbf{g}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \iff \mathbf{G}^T\mathbf{G} = \mathbf{I}$$

- Isso implica:

$$\mathbf{z}^{(i)} = \mathbf{G}^T\mathbf{x}^{(i)} \qquad \text{e} \qquad J = \frac{1}{m}\sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \mathbf{G}\mathbf{G}^T\mathbf{x}^{(i)}\|^2$$

Formulação Matemática

- Note que

$$
\begin{aligned}
\|\mathbf{x}^{(i)} - \mathbf{G}\mathbf{G}^T\mathbf{x}^{(i)}\|^2 &= \|(\mathbf{I} - \mathbf{G}\mathbf{G}^T)\mathbf{x}^{(i)}\|^2 \\
&= {\mathbf{x}^{(i)}}^T(\mathbf{I} - \mathbf{G}\mathbf{G}^T)^T(\mathbf{I} - \mathbf{G}\mathbf{G}^T)\mathbf{x}^{(i)} \\
&= {\mathbf{x}^{(i)}}^T(\mathbf{I} - 2\mathbf{G}\mathbf{G}^T + \mathbf{G}\mathbf{G}^T\mathbf{G}\mathbf{G}^T)\mathbf{x}^{(i)} \\
&= {\mathbf{x}^{(i)}}^T(\mathbf{I} - \mathbf{G}\mathbf{G}^T)\mathbf{x}^{(i)} \\
&= \|\mathbf{x}^{(i)}\|^2 - {\mathbf{x}^{(i)}}^T\mathbf{G}\mathbf{G}^T\mathbf{x}^{(i)} \\
&= \|\mathbf{x}^{(i)}\|^2 - \|\mathbf{z}^{(i)}\|^2
\end{aligned}
$$

Formulação Matemática

- Desejamos minimizar

$$J = \frac{1}{m}\sum_{i=1}^{m} \|\mathbf{x}^{(i)}\|^2 - \frac{1}{m}\sum_{i=1}^{m} \|\mathbf{z}^{(i)}\|^2$$

o que equivale a maximizar

$$\frac{1}{m}\sum_{i=1}^{m} \|\mathbf{z}^{(i)}\|^2 = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} (z_k^{(i)})^2 = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} (\mathbf{g}_k^T \mathbf{x}^{(i)})^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{g}_k^T \mathbf{x}^{(i)} \mathbf{x}^{(i)^T} \mathbf{g}_k = \sum_{k=1}^{K} \mathbf{g}_k^T \mathbf{C} \mathbf{g}_k$$

onde

$$\mathbf{C} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}^{(i)} \mathbf{x}^{(i)^T} = \frac{1}{m}\mathbf{X}^T\mathbf{X}$$

é a matriz de covariância amostral de $\mathbf{x}$

Formulação Matemática

- A solução ótima para a maximização de

$$\sum_{k=1}^{K} \mathbf{g}_k^T \mathbf{C} \mathbf{g}_k = \sum_{k=1}^{K} \mathbf{g}_k^T \left(\frac{1}{m} \mathbf{X}^T \mathbf{X} \right) \mathbf{g}_k$$

é dada pelos autovetores de $\mathbf{C}$ associados aos K maiores autovalores, obtidos pela decomposição

$$\mathbf{C} = \mathbf{V} \Lambda \mathbf{V}^T$$

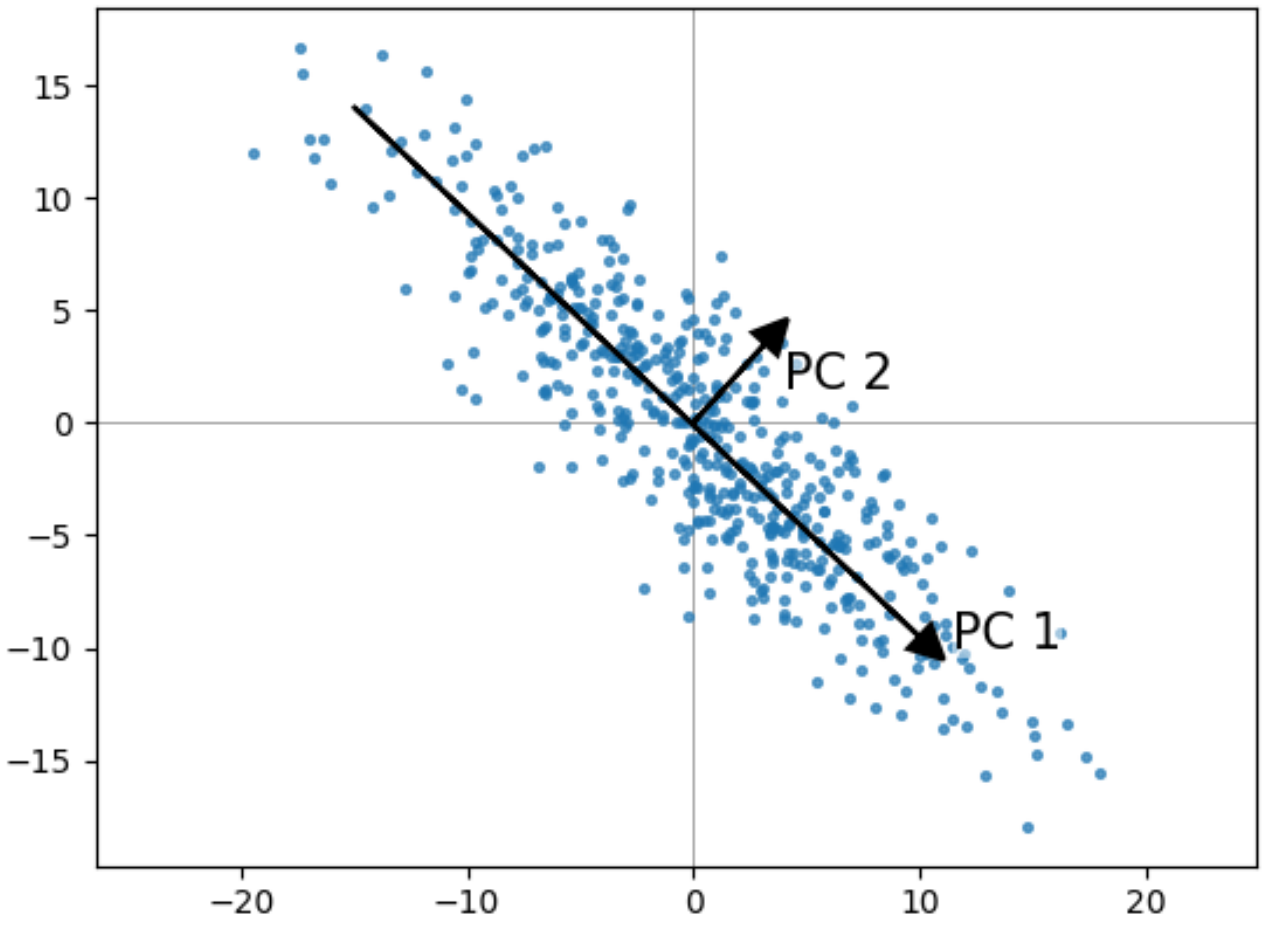
onde $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1 \geq \cdots \geq \lambda_n$ e $\mathbf{V}^T \mathbf{V} = I$.

- Extrai-se as K primeiras colunas de $\mathbf{V}$:

$$\mathbf{G} = \mathbf{V}[:, :K] = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_K \\ | & & | \end{bmatrix}$$

os quais são chamados de K componentes principais

Exemplo



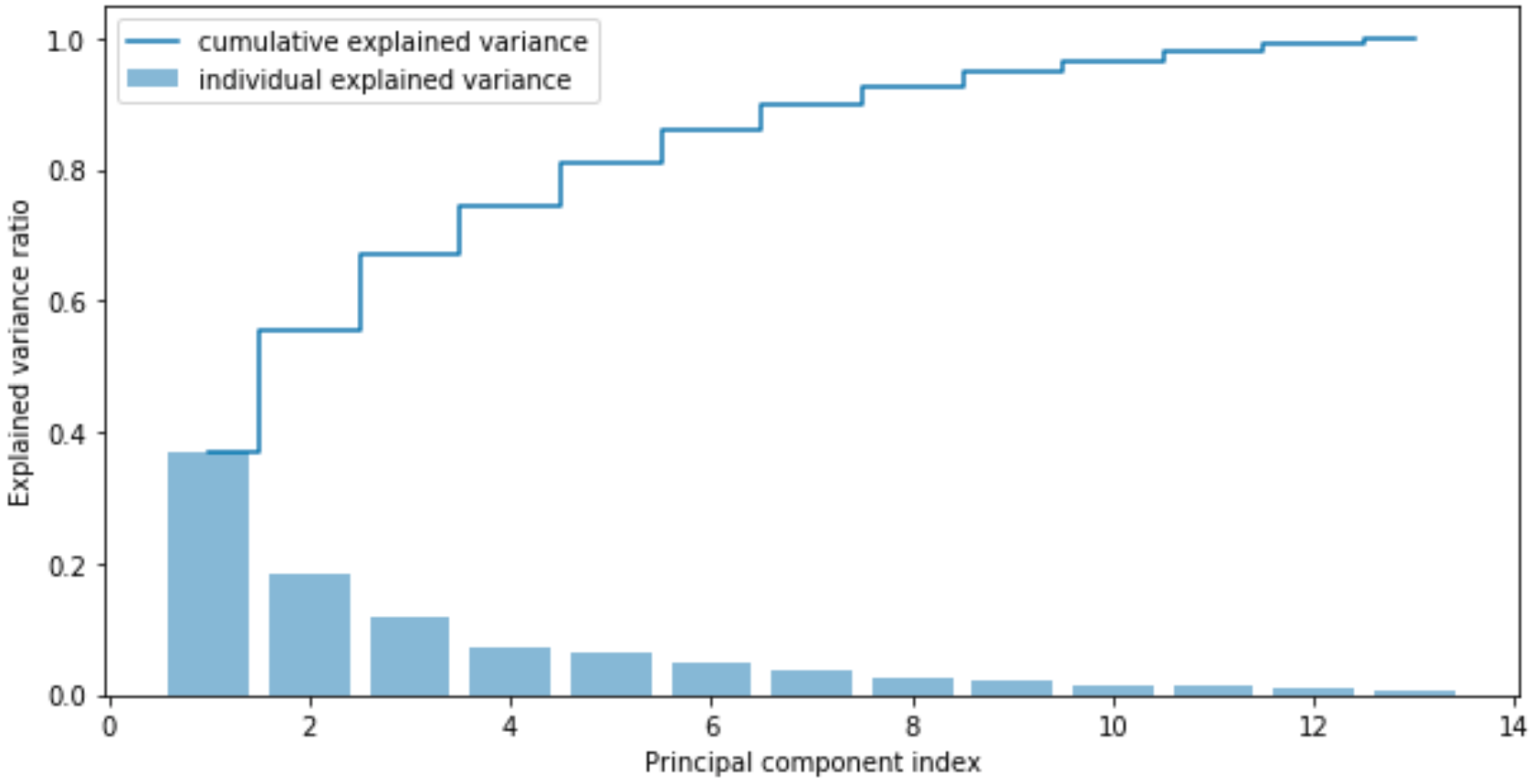
Interpretação por Maximização da Variância

- Note que o custo pode ser interpretado como

$$
\begin{aligned}
J &= \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}^{(i)}\|^2 \quad - \quad \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{z}^{(i)}\|^2 \\
&= \sum_{j=1}^{n}\frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)})^2 \quad - \quad \sum_{k=1}^{K}\frac{1}{m}\sum_{i=1}^{m}(z_k^{(i)})^2 \\
&= \underbrace{\sum_{j=1}^{n}\lambda_j}_{\text{variância total}} \quad - \quad \underbrace{\sum_{k=1}^{K}\lambda_k}_{\text{variância retida/“explicada”}} \quad = \sum_{j=K+1}^{n}\lambda_j
\end{aligned}
$$

- Portanto, o problema equivale a maximizar a variância da projeção
- Percentual de variância retida: $(\sum_{j=1}^{K}\lambda_j)/(\sum_{j=1}^{n}\lambda_j)$

Exemplo



- Uma regra prática é escolher K tal que o percentual de variância explicada seja maior que um dado valor (ex: 90%)

Decomposição em Valores Singulares

- Uma forma prática de obter os autovetores de $\frac{1}{m}\mathbf{X}^T\mathbf{X}$ é através da decomposição em valores singulares (SVD) de $\mathbf{X}$, dada por:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$$

  - $\mathbf{U} \in \mathbb{R}^{m\times m}$ satisfaz $\mathbf{U}^T\mathbf{U} = \mathbf{I}$
  - $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{m\times n},\ \sigma_1, \ldots, \sigma_r \geq 0,\ r = \min(m, n)$
  - $\mathbf{V} \in \mathbb{R}^{n\times n}$ satisfaz $\mathbf{V}^T\mathbf{V} = \mathbf{I}$

- Consequentemente,

$$\frac{1}{m}\mathbf{X}^T\mathbf{X} = \frac{1}{m}\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T = \frac{1}{m}\mathbf{V}\Sigma^T\Sigma\mathbf{V}^T = \mathbf{V}\Lambda\mathbf{V}^T$$

onde $\lambda_j = \sigma_j^2/m,\ j = 1, \ldots, n$

Análise de Componentes Principais

Principal Component Analysis (PCA)

- Escolha $K \leq n$
- Treinamento (assumindo $\mathbf{X}$ centralizado):

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T \qquad \text{(SVD)}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_K \end{bmatrix} = \mathbf{V}\texttt{[:,:K]} \qquad \text{(componentes principais)}$$

  onde $\lambda_k = \sigma_k^2/m$ é o autovalor associado a $\mathbf{v}_k$, $k = 1, \ldots, K$

- Transformação (codificação):

$$\mathbf{z} = \mathbf{G}^T\mathbf{x}$$

- Reconstrução (decodificação):

$$\hat{\mathbf{x}} = \mathbf{G}\mathbf{z}$$

Análise de Componentes Principais

Principal Component Analysis (PCA)

- Escolha $K \leq n$
- Treinamento:

$$\mathbf{X}' = \mathbf{X} - \boldsymbol{\mu}^T, \quad \boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}^{(i)} \qquad \text{(centralização)}$$

$$\mathbf{X}' = \mathbf{U}\Sigma\mathbf{V}^T \qquad \text{(SVD)}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_K \end{bmatrix} = \mathbf{V}\texttt{[:,:K]} \qquad \text{(componentes principais)}$$

  onde $\lambda_k = \sigma_k^2/m$ é o autovalor associado a $\mathbf{v}_k$, $k = 1, \ldots, K$

- Transformação (codificação):

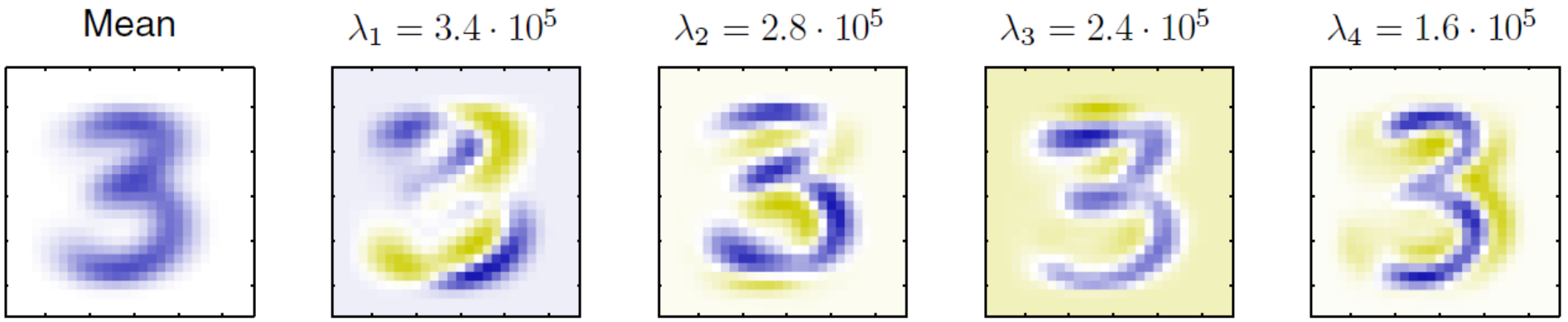
$$\mathbf{z} = \mathbf{G}^T(\mathbf{x} - \boldsymbol{\mu})$$

- Reconstrução (decodificação):

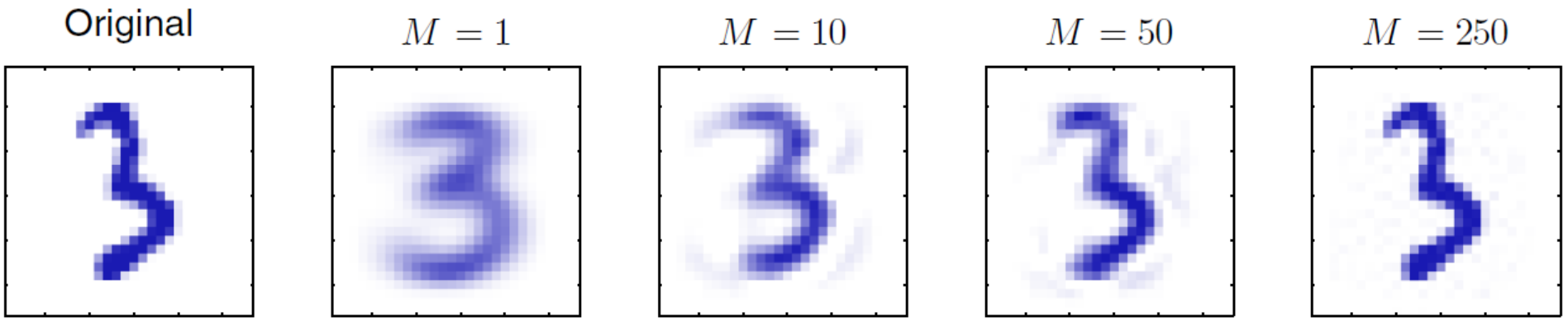
$$\hat{\mathbf{x}} = \boldsymbol{\mu} + \mathbf{G}\mathbf{z}$$

Exemplo

- Média e autovetores (positivo/negativo):



- Reconstrução ($M = K$):



Exemplo: *Eigenfaces* (1)

Amostras:

Exemplo: *Eigenfaces* (1)

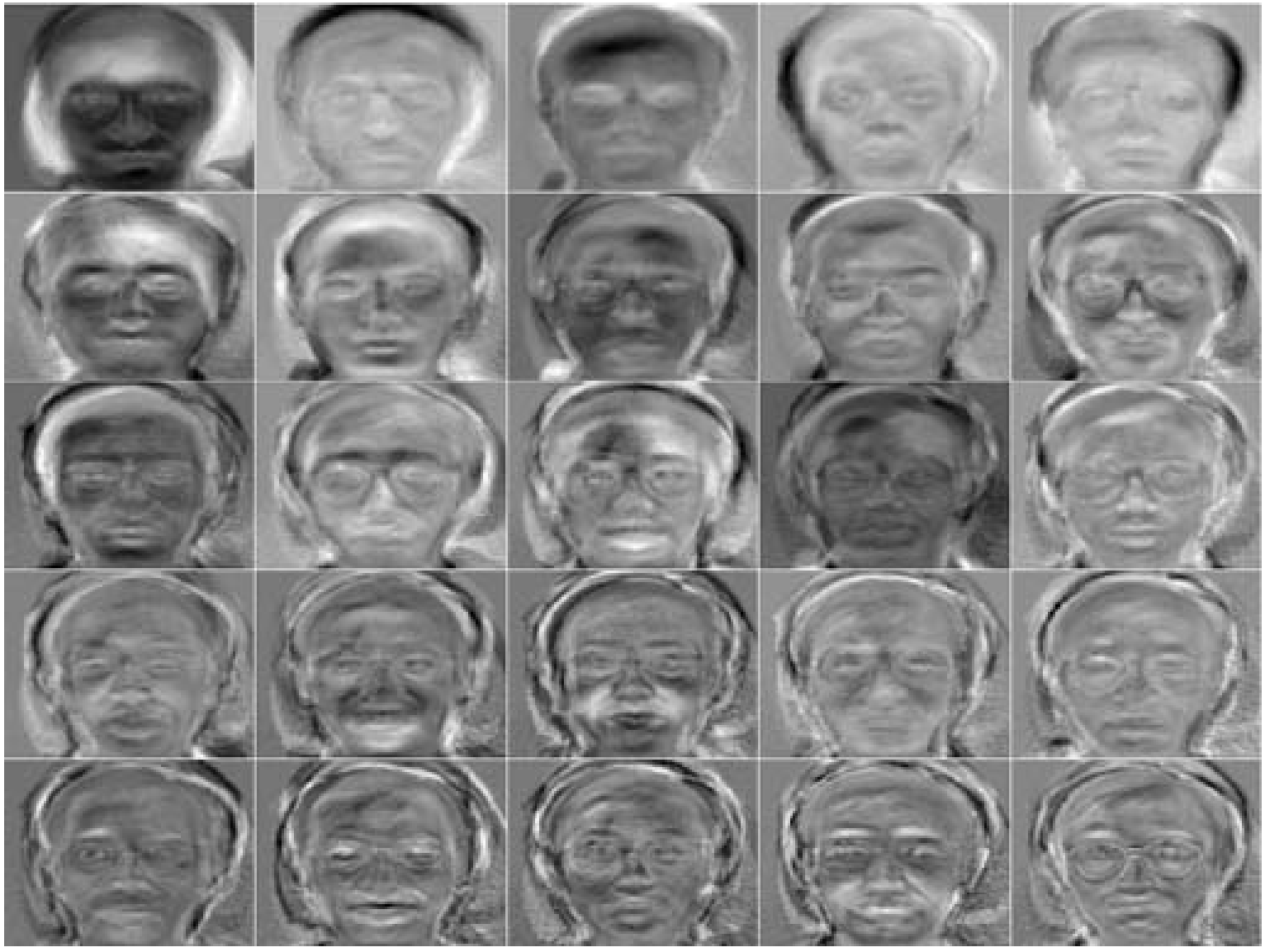
Média e componentes principais:

Exemplo: *Eigenfaces* (1)

Original e reconstrução:

Exemplo: *Eigenfaces* (2)

Componentes principais:



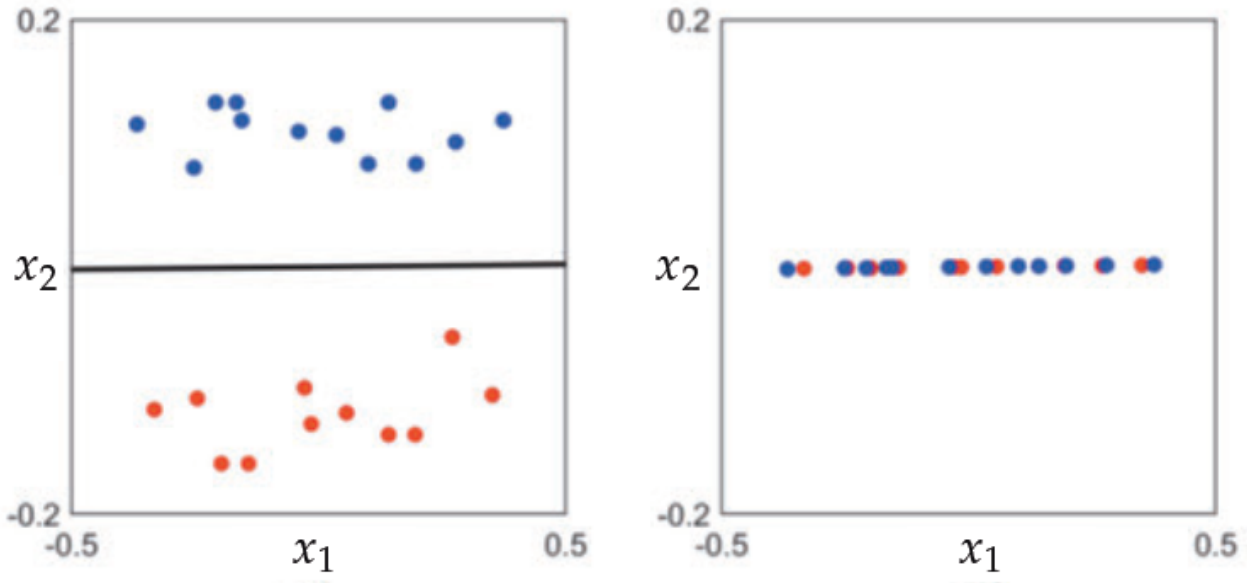
Exemplo: *Eigenfaces* (2)

Média e reconstrução:

Uso em Aprendizado Supervisionado

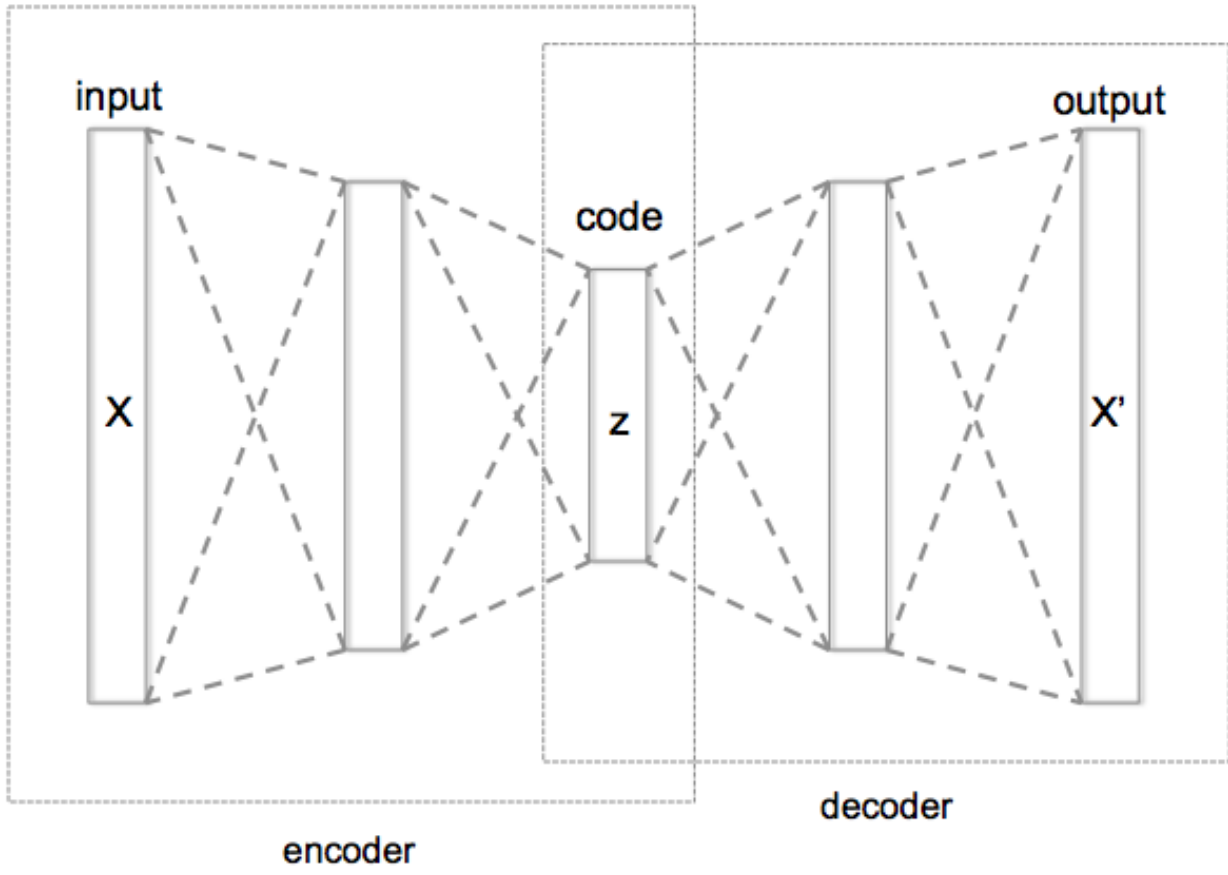
- Quando usado como pré-processamento em aprendizado supervisionado, os parâmetros $\boldsymbol{\mu}$ e $\mathbf{G}$ devem ser estimados no conjunto de treinamento e aplicados (sem alteração) no conjunto de teste

- PCA é um método não-supervisionado: ignora rótulos $y^{(i)}$
  - No contexto da classificação, a direção de máxima variância não necessariamente fornece a melhor separação entre classes



Redes Neurais *Autoencoders*

Redes Neurais *Autoencoder*



- $\mathbf{z} = f(\mathbf{x})$ e $\hat{\mathbf{x}} = g(\mathbf{z})$
- Treinada com erro de reconstrução (ex: quadrático):

$$L(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$$