Lecture 8
Principal Component Analysis

Luigi Freda

ALCOR Lab
DIAG
University of Rome "La Sapienza"

December 13, 2016

Outline

Outline

Eigenvalues

given a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$

- a nonzero vector $\mathbf{v} \in \mathbb{C}^n$ is said to be its (right) **eigenvector** if

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

  for some scalar $\lambda \in \mathbb{C}$; λ is called an **eigenvalue** of $\mathbf{A}$
- the set of all eigenvalues of $\mathbf{A}$ is called its **spectrum** and it's denoted by $\sigma(\mathbf{A})$
- the Matlab command

  $[\mathrm{V}, \mathrm{D}] = \ \mathbf{eig}\,(\mathrm{A})$

  produces a diagonal matrix $\mathbf{D}$ of eigenvalues and a matrix $\mathbf{V}$ such that

$$\mathbf{AV} = \mathbf{VD} \qquad \text{with } \mathbf{D} = \mathrm{diag}(\lambda_i)$$

- if $\mathbf{A}$ is **diagonalizable** then $\mathbf{V}$ is full-rank (i.e. $\mathrm{rank}(V) = n$) and the columns of $\mathbf{V}$ correspond to n linearly independent eigenvectors, i.e. $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$; in this case one has

$$\mathbf{A} = \mathbf{VDV}^{-1} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

- if $\tilde{\mathbf{A}} = \mathbf{TAT}^{-1}$, where $\mathbf{T}$ is a nonsingular matrix, then $\tilde{\mathbf{A}}$ and $\mathbf{A}$ are called **similar matrices**; at the previous point, $\mathbf{A}$ and $\mathbf{D}$ are similar matrices

Outline

Eigenvalues

Real Matrix

given a **real matrix** $\mathbf{A} \in \mathbb{R}^{n \times n}$

- all its eigenvalues $\sigma(\mathbf{A})$ are the roots of the **characteristic polynomial equation**

$$\det(\lambda \mathbf{I} - \mathbf{A}) = 0$$

- given that $\mathbf{A}$ is a real matrix, if $\lambda^* \in \mathbb{C}$ is an eigenvalue then its **conjugate** $\overline{\lambda^*} \in \mathbb{C}$ is also an eigenvalue, i.e., $\sigma(\mathbf{A}) = \overline{\sigma}(\mathbf{A})$

$$\det(\lambda^* \mathbf{I} - \mathbf{A}) = 0 \implies \overline{\det(\lambda^* \mathbf{I} - \mathbf{A})} = \det(\overline{\lambda^*} \mathbf{I} - \overline{\mathbf{A}}) = \det(\overline{\lambda^*} \mathbf{I} - \mathbf{A}) = 0$$

- $\sigma(\mathbf{A}) = \sigma(\mathbf{A}^T)$ since

$$\det(\lambda \mathbf{I} - \mathbf{A}) = \det((\lambda \mathbf{I} - \mathbf{A})^T) = \det(\lambda \mathbf{I} - \mathbf{A}^T)$$

- if $\tilde{\mathbf{A}} = \mathbf{T} \mathbf{A} \mathbf{T}^{-1}$, where $\mathbf{T}$ is a nonsingular matrix, then $\sigma(\tilde{\mathbf{A}}) = \sigma(\mathbf{A})$ since

$$\det(\lambda \mathbf{I} - \tilde{\mathbf{A}}) = \det(\lambda \mathbf{T} \mathbf{T}^{-1} - \mathbf{T} \mathbf{A} \mathbf{T}^{-1}) = \det(\mathbf{T}(\lambda \mathbf{I} - \mathbf{A}) \mathbf{T}^{-1}) =$$

$$= \det(\mathbf{T}) \det(\lambda \mathbf{I} - \mathbf{A}) \det(\mathbf{T}^{-1}) = \det(\lambda \mathbf{I} - \mathbf{A}) \qquad \left(\det \mathbf{I} = \det(\mathbf{T}^{-1}) \det \mathbf{T} = 1\right)$$

given a **real matrix** $\mathbf{A} \in \mathbb{R}^{n \times n}$

- eigenvectors of $\mathbf{A}$ associated to different eigenvalues are **linearly independent**

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \lambda_1 \neq \lambda_2, \quad \mathbf{v}_1 \neq \mathbf{v}_2$$

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 = 0 \implies \mathbf{A}(\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2) = 0 \implies \alpha_1 \lambda_1 \mathbf{v}_1 + \alpha_2 \lambda_2 \mathbf{v}_2 = 0$$

since $\lambda_1(\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2) = 0 \implies \alpha_2(\lambda_2 - \lambda_1)\mathbf{v}_2 = 0 \implies \alpha_2 = 0 \implies \alpha_1 = 0$

- if $|\sigma(\mathbf{A})| = n$, i.e. $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ with $\lambda_i \neq \lambda_j$ for $i \neq j$, then $\mathbf{A}$ is **diagonalizable**

Eigenvalues

Real Matrix

given a **real matrix** $\mathbf{A} \in \mathbb{R}^{n \times n}$

- in general $\sigma(\mathbf{A}) = \{\lambda_1, ..., \lambda_m\}$ with $m \leq n$ and $\lambda_i \neq \lambda_j$ for $i \neq j$, and one has

$$\det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - \lambda_1)^{\mu_a(\lambda_1)}(\lambda - \lambda_2)^{\mu_a(\lambda_2)}...(\lambda - \lambda_m)^{\mu_a(\lambda_m)} \qquad \sum_{i=1}^{m} \mu_a(\lambda_i) = n$$

  where $\mu_a(\lambda_i)$ is the **algebraic multiplicity** of λ_i

- in the general case one can always find the **Jordan canonical form**

$$\mathbf{A} = \mathbf{V}\mathbf{J}\mathbf{V}^{-1} \qquad \mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_p \end{bmatrix} \qquad \mathbf{J_i} = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$$

Eigenvalues

Real Matrix

given a **real matrix** $\mathbf{A} \in \mathbb{R}^{n \times n}$

- let $E(\lambda_i) \triangleq \ker(\mathbf{A} - \lambda_i \mathbf{I}) = \{\mathbf{v} \in \mathbb{R}^n : (\mathbf{A} - \lambda_i \mathbf{I})\mathbf{v} = 0\}$ be the **eigenspace** of λ_i
- let $\mu_g(\lambda_i) \triangleq \dim(E(\lambda_i))$ be the **geometric multiplicity** of λ_i $\quad (\mu_g(\lambda_i) \leq \mu_a(\lambda_i))$
- if $E(\lambda_i) = \text{span}\{\mathbf{v}_{i1}, ..., \mathbf{v}_{i\mu_i}\}$, with $\mu_i \triangleq \mu_g(\lambda_i)$, then

$$\mathbf{A}\mathbf{V}_i = \mathbf{V}_i\mathbf{D}_i \qquad \text{with } \mathbf{V}_i = [\mathbf{v}_{i1}, ..., \mathbf{v}_{i\mu_i}] \in \mathbb{R}^{n \times \mu_i} \text{ and } \mathbf{D}_i = \lambda_i \mathbf{I}_{\mu_i}$$

- if $\mu_g(\lambda_i) = \mu_a(\lambda_i)$ for each $i \in \{1, ..., m\}$ then

$$\mathbb{R}^n = E(\lambda_1) \oplus ... \oplus E(\lambda_m) \qquad \sum_{i=1}^{n} \mu_g(\lambda_i) = n$$

and **A** is **diagonalizable**

Eigenvalues

Real Symmetric Matrix

given a **real symmetric matrix** $\mathbf{S} \in \mathbb{R}^{n \times n}$ (i.e. $\mathbf{S} = \mathbf{S}^T$)

- all eigenvalues of $\mathbf{S}$ are real, i.e. $\sigma(\mathbf{S}) \subset \mathbb{R}$
- eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ of $\mathbf{S}$ associated to different eigenvalues are **linearly independent and orthogonal**

$$\mathbf{S}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \lambda_1 \neq \lambda_2, \quad \mathbf{v}_1 \neq \mathbf{v}_2, \qquad \mathbf{v}_2^T(\mathbf{S}\mathbf{v}_1 - \lambda_1 \mathbf{v}_1) = 0$$

$$\implies (\mathbf{S}\mathbf{v}_2)^T \mathbf{v}_1 - \lambda_1 \mathbf{v}_2^T \mathbf{v}_1 = 0 \implies \mathbf{v}_2^T \mathbf{v}_1 (\lambda_2 - \lambda_1) = 0 \implies \mathbf{v}_2^T \mathbf{v}_1 = 0$$

- $\mathbf{S}$ is **diagonalizable** and one can find an **orthonormal basis** $\mathbf{V}$ such that $\mathbf{V}^T = \mathbf{V}^{-1}$

$$\mathbf{S} = \mathbf{V}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

- if $\mathbf{S} > 0$ ($\mathbf{S} \geq 0$) then $\lambda_i > 0$ ($\lambda_i \geq 0$) for $i \in \{1, ..., m\}$

N.B.: the above results can be applied for instance to covariance matrices $\mathbf{\Sigma}$ which are symmetric and positive definite

Outline

Singular Value Decomposition

idea: the singular values generalize the notion of eigenvalues to any kind of matrix

given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$

- this can be decomposed as follows

$$\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\mathbf{U}}_{N \times N} \underbrace{\mathbf{S}}_{N \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

- $\mathbf{U} \in \mathbb{R}^{N \times N}$ is orthonormal, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_N$
- $\mathbf{V} \in \mathbb{R}^{D \times D}$ is orthonormal, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_D$
- $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0$ are the **singular values**, $r = \min(N, D)$ and $\mathbf{S} \in \mathbb{R}^{N \times D}$

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_D \\ \hline & \mathbf{0}_{N-D} & \end{bmatrix} \text{ if } N > D \quad \text{or} \quad \mathbf{S} = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0}_{D-N} \\ & & \sigma_N & \end{array}\right] \text{ if } N \leq D$$

Outline

Singular Value Decomposition

Connection with Eigenvalues

given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$

- one has $\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\mathbf{U}}_{N \times N} \underbrace{\mathbf{S}}_{N \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ where $r = \min(N, D)$
- $\mathbf{U} \in \mathbb{R}^{N \times N}$ is orthonormal, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_N$
- $\mathbf{V} \in \mathbb{R}^{D \times D}$ is orthonormal, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_D$
- $\mathbf{S} \in \mathbb{R}^{N \times D}$ contains the singular values σ_i on the main diagonal and 0s elsewhere

- we have

$$\mathbf{X}^T\mathbf{X} = (\mathbf{V}\mathbf{S}^T\mathbf{U}^T)(\mathbf{U}\mathbf{S}\mathbf{V}^T) = \mathbf{V}\mathbf{S}^T\mathbf{S}\mathbf{V}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T$$

  where $\mathbf{D} \triangleq \mathbf{S}^T\mathbf{S}$ is a diagonal matrix containing the squared singular values σ_i^2 (and possibly some additional zeros) on the main diagonal and 0s elsewhere
- since $(\mathbf{X}^T\mathbf{X})\mathbf{V} = \mathbf{V}\mathbf{D}$, then $\mathbf{V}$ contains the eigenvectors of $\mathbf{X}^T\mathbf{X}$
- the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are the squared singular values σ_i^2, contained in $\mathbf{D}$
- the columns of $\mathbf{V}$ are called the **right singular vectors** of $\mathbf{X}$

Singular Value Decomposition

Connection with Eigenvalues

given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$

- one has $\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\mathbf{U}}_{N \times N} \underbrace{\mathbf{S}}_{N \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ where $r = \min(N, D)$
- $\mathbf{U} \in \mathbb{R}^{N \times N}$ is orthonormal, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_N$
- $\mathbf{V} \in \mathbb{R}^{D \times D}$ is orthonormal, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_D$
- $\mathbf{S} \in \mathbb{R}^{N \times D}$ contains the singular values σ_i on the main diagonal and 0s elsewhere

- we have
$$\mathbf{X}\mathbf{X}^T = (\mathbf{U}\mathbf{S}\mathbf{V}^T)(\mathbf{V}\mathbf{S}^T\mathbf{U}^T) = \mathbf{U}\mathbf{S}\mathbf{S}^T\mathbf{U}^T = \mathbf{U}\hat{\mathbf{D}}\mathbf{U}^T$$
where $\hat{\mathbf{D}} \triangleq \mathbf{S}\mathbf{S}^T$ is a diagonal matrix containing the squared singular values σ_i^2 (and possibly some additional zeros) on the main diagonal and 0s elsewhere
- since $(\mathbf{X}\mathbf{X}^T)\mathbf{U} = \mathbf{U}\hat{\mathbf{D}}$, then $\mathbf{U}$ contains the eigenvectors of $\mathbf{X}\mathbf{X}^T$
- the eigenvalues of $\mathbf{X}\mathbf{X}^T$ are the squared singular values σ_i^2, contained in $\hat{\mathbf{D}}$
- the columns of $\mathbf{U}$ are called the **left singular vectors** of $\mathbf{X}$

Outline

Singular Value Decomposition

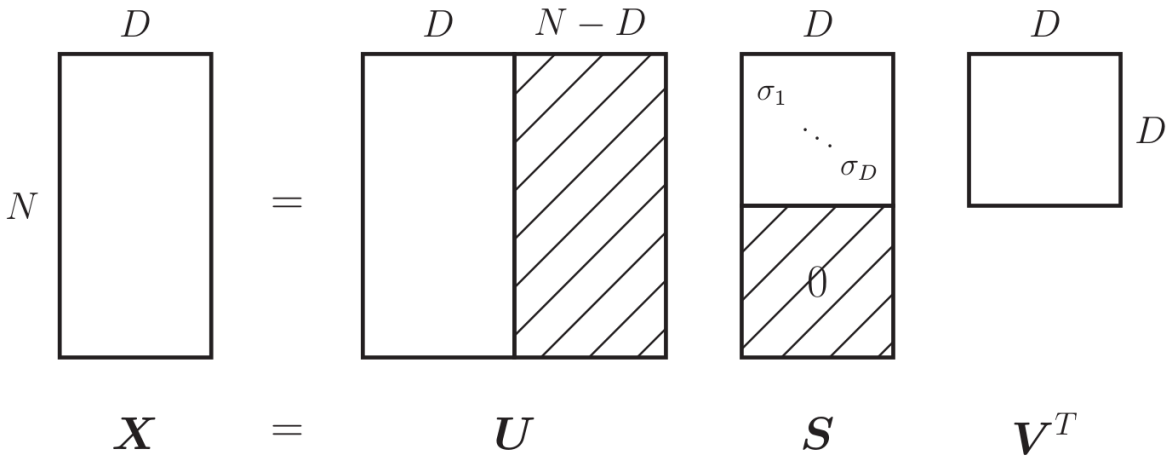
given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ with $N > D$

- this can be decomposed as follows

$$\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\mathbf{U}}_{N \times N} \underbrace{\mathbf{S}}_{N \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{D} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_D \\ \hline & \mathbf{0}_{N-D} & \end{bmatrix}$$

- the last $N - D$ columns of $\mathbf{U}$ are irrelevant, since they will be multiplied by zero



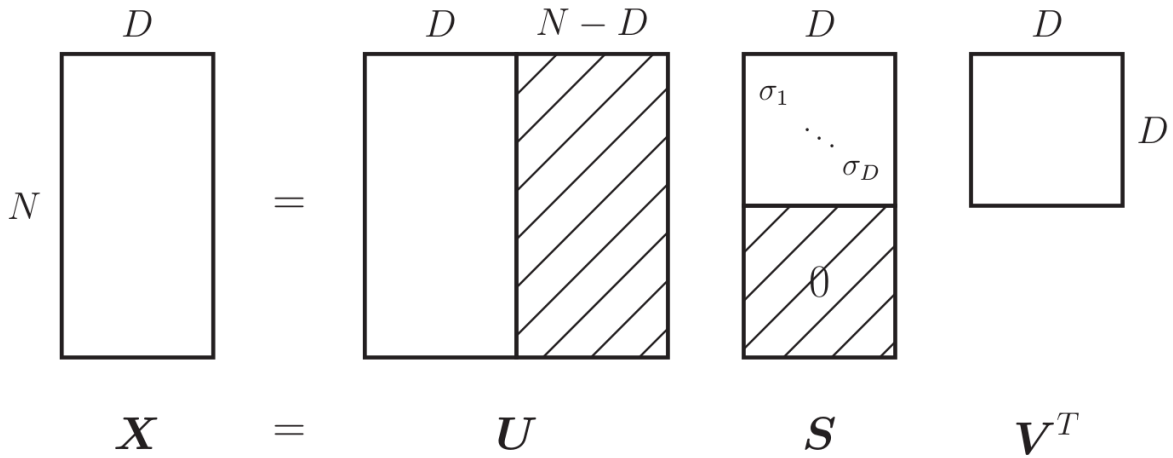
Singular Value Decomposition

given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ with $N > D$

- this can be decomposed as $\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\mathbf{U}}_{N \times N} \underbrace{\mathbf{S}}_{N \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{D} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
- since the last $N - D$ columns of $\mathbf{U}$ are irrelevant, we can directly neglect them and consider the **economy sized SVD**

$$\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\hat{\mathbf{U}}}_{N \times D} \underbrace{\hat{\mathbf{S}}}_{D \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{D} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

where $\hat{\mathbf{U}}^T \hat{\mathbf{U}} = \mathbf{I}_D$, $\mathbf{V}^T \mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_D$ and $\hat{\mathbf{S}} = \text{diag}(\sigma_1, ..., \sigma_D)$



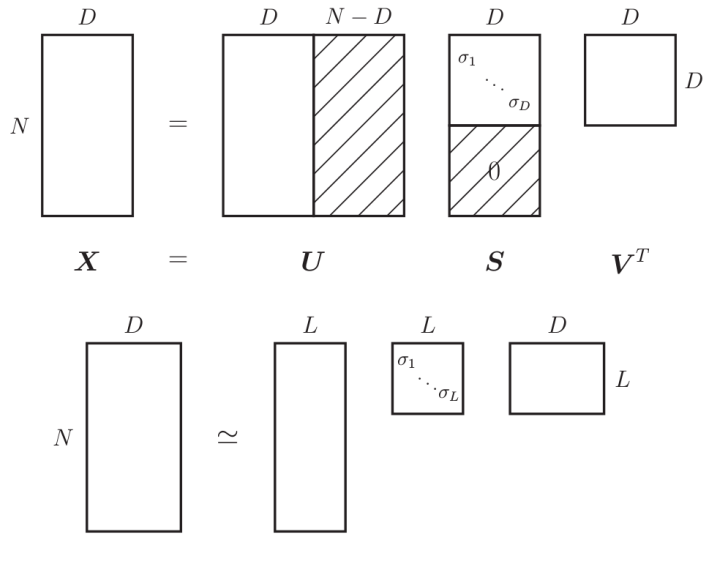
Singular Value Decomposition

given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ with $N > D$

- this can be decomposed as $\underbrace{\mathbf{X}}_{N \times D} = \underbrace{\mathbf{U}}_{N \times N} \underbrace{\mathbf{S}}_{N \times D} \underbrace{\mathbf{V}^T}_{D \times D} = \sum_{i=1}^{D} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
- a **rank L approximation** of $\mathbf{X}$ considers only the first $L < D$ columns of $\mathbf{U}$ and the first L singular values
- we can then consider the **truncated SVD**

$$\underbrace{\mathbf{X}_L}_{N \times D} = \underbrace{\mathbf{U}_L}_{N \times L} \underbrace{\mathbf{S}_L}_{L \times L} \underbrace{\mathbf{V}_L^T}_{D \times L} = \sum_{i=1}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

where $\mathbf{U}_L^T \mathbf{U}_L = \mathbf{I}_L$, $\mathbf{V}_L^T \mathbf{V}_L = \mathbf{I}_D$ and $\mathbf{S}_L = \text{diag}(\sigma_1, ..., \sigma_L)$



Singular Value Decomposition

given a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ with $N > D$

- it is possible to show that $\mathbf{X}_L = \mathbf{U}_L \mathbf{S}_L \mathbf{V}_L^T$ is the **best rank L approximation** matrix which solve

$$\mathbf{X}_L = \underset{\hat{\mathbf{X}}}{\operatorname{argmin}} \|\mathbf{X} - \hat{\mathbf{X}}\|_F \qquad \text{with the constraint} \quad \operatorname{rank}(\hat{\mathbf{X}}) = L$$

  where $\|\mathbf{A}\|_F$ denotes the **Frobenious norm**

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} = \left(\operatorname{trace}(\mathbf{A}^T \mathbf{A}) \right)^{1/2}$$

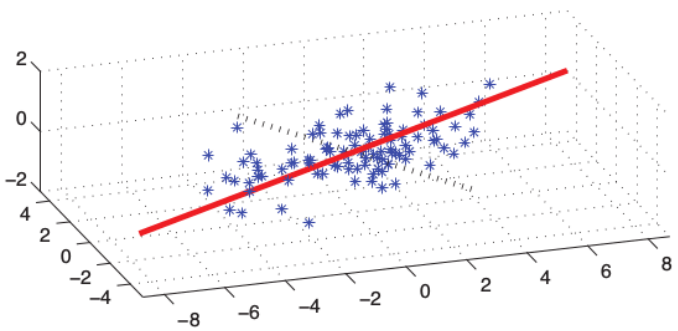
- by replacing $\hat{\mathbf{X}} = \mathbf{X}_L$ one obtains

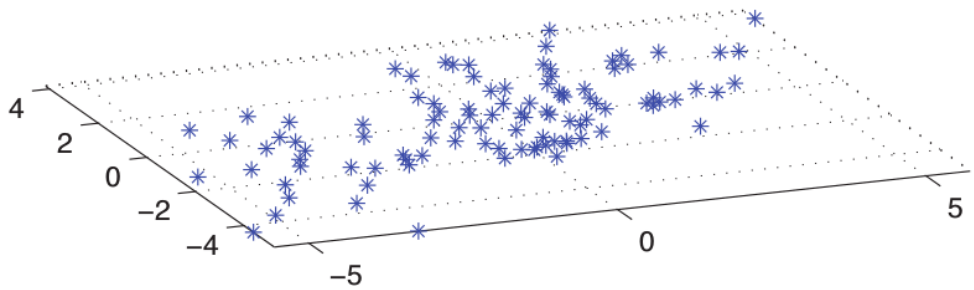
$$\|\mathbf{X} - \mathbf{X}_L\|_F = \|\mathbf{U}(\mathbf{S} - \begin{bmatrix} \mathbf{S}_L & \\ & \mathbf{0} \end{bmatrix})\mathbf{V}^T\|_F = \|\mathbf{S} - \begin{bmatrix} \mathbf{S}_L & \\ & \mathbf{0} \end{bmatrix}\|_F = \sum_{i=L+1}^{r} \sigma_i$$

Outline

Discovering Latent Factors

An example

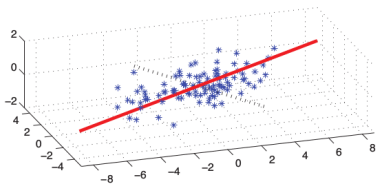


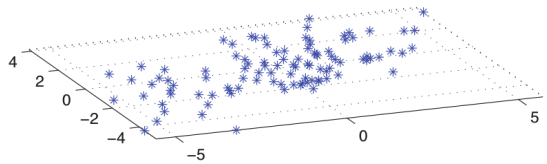


- **dimensionality reduction**: it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the "essence" of the data
- for example in the above plot:
  - the 2d approximation is quite good, most points lie close to this subspace
  - projecting points onto the red line 1d approx is a rather poor approximation

Discovering Latent Factors

Motivations





- although data may appear high dimensional, there may only be a small number of degrees of variability, corresponding to **latent factors**
- low dimensional representations, when used as input to statistical models, often result in **better predictive accuracy** (focusing on the "essence")
- low dimensional representations are useful for **enabling fast** nearest neighbor **searches**
- two dimensional projections are very useful for **visualizing high dimensional data**

Principal Component Analysis

PCA

- the most common approach to dimensionality reduction is called **Principal Components Analysis** or **PCA**
- given $\mathbf{x}_i \in \mathbb{R}^D$, we compute an approximation $\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i$ with
  - $\mathbf{z}_i \in \mathbb{R}^L$ where $L < D$ (dimensionality reduction)
  - $\mathbf{W} \in \mathbb{R}^{D \times L}$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ (orthogonal $\mathbf{W}$)

  so as to minimize the **reconstruction error**

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{W}\mathbf{z}_i\|^2$$

- since $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ one has $\|\mathbf{x}\|_2 = (\mathbf{x}^T\mathbf{x})^{1/2} = (\mathbf{z}^T\mathbf{W}^T\mathbf{W}\mathbf{z})^{1/2} = (\mathbf{z}^T\mathbf{z})^{1/2} = \|\mathbf{z}\|_2$
- in 3D or 2D $\mathbf{W}$ can represent part of a rotation matrix
- this can be thought of as an unsupervised version of (multi-output) linear regression, where we observe the high-dimensional response $\mathbf{y} = \mathbf{x}$, but not the low-dimensional "cause" $\mathbf{z}$

- the objective function (reconstruction error)

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{W}\mathbf{z}_i\|^2$$

is equivalent to

$$J(\mathbf{W}, \mathbf{Z}) = \|\mathbf{X}^T - \mathbf{W}\mathbf{Z}^T\|_F^2$$

where $\|\mathbf{A}\|_F$ denotes the **Frobenious norm**

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \left(\text{trace}(\mathbf{A}^T\mathbf{A})\right)^{1/2}$$

Principal Component Analysis

PCA

- we have to minimize the objective function

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{W}\mathbf{z}_i\|^2$$

  or equivalently

$$J(\mathbf{W}, \mathbf{Z}) = \|\mathbf{X}^T - \mathbf{W}\mathbf{Z}^T\|_F^2$$

  subject to $\mathbf{W} \in \mathbb{R}^{D\times L}$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

- it is possible to prove that the **optimal solution** is obtained by setting

$$\hat{\mathbf{W}} = \mathbf{V}_L$$

$$\hat{\mathbf{Z}} = \mathbf{X}\mathbf{W} \qquad (\mathbf{z}_i = \hat{\mathbf{W}}^T\mathbf{x}_i = \mathbf{V}_L^T\mathbf{x}_i)$$

  where $\mathbf{X}_L = \mathbf{U}_L\mathbf{S}_L\mathbf{V}_L^T$ is the rank L truncated SVD of $\mathbf{X}$

- we have $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$ hence $(\mathbf{X}^T\mathbf{X})\mathbf{V}_L = \mathbf{V}_L\mathbf{S}_L^2$ since $\mathbf{V} = [\mathbf{V}_L|*]$
- $\mathbf{V}_L$ contains the L eigenvectors (**principal directions**) with the **largest eigenvalues** of the empirical covariance matrix $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^T\mathbf{X} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$
- $\mathbf{z}_i = \mathbf{V}_L^T\mathbf{x}_i$ is the orthogonal projection of $\mathbf{x}_i$ on the eigenvectors in $\mathbf{V}_L$

- by replacing $\mathbf{Z} = \mathbf{XW}$ in the objective function and recalling that $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

$$J(\mathbf{Z}) = \|\mathbf{X}^T - \mathbf{WZ}^T\|_F^2 = \text{trace}\Big((\mathbf{X}^T - \mathbf{WZ}^T)^T(\mathbf{X}^T - \mathbf{WZ}^T)\Big) =$$

$$= \text{trace}(\mathbf{XX}^T - \mathbf{XWZ}^T - \mathbf{ZW}^T\mathbf{X}^T + \mathbf{ZZ}^T) = \text{trace}(\mathbf{X}^T\mathbf{X}) - \text{trace}(\mathbf{W}^T\mathbf{X}^T\mathbf{XW}) =$$

$$= \text{trace}(\mathbf{X}^T\mathbf{X}) - \text{trace}(\mathbf{Z}^T\mathbf{Z}) = \text{trace}(N\Sigma_X) - \text{trace}(N\Sigma_Z)$$

  where we used the following facts: $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$, $\text{trace}(\mathbf{A}^T) = \text{trace}(\mathbf{A})$

- **minimizing the reconstruction error** is equivalent to **maximizing the variance of the projected data $\mathbf{z}_i = \mathbf{W}\mathbf{x}_i$**

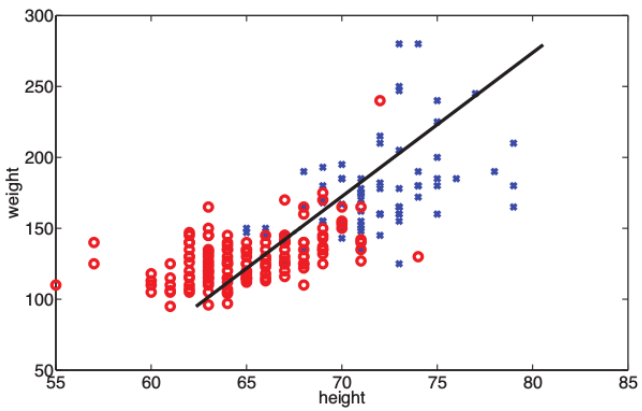
Principal Component Analysis

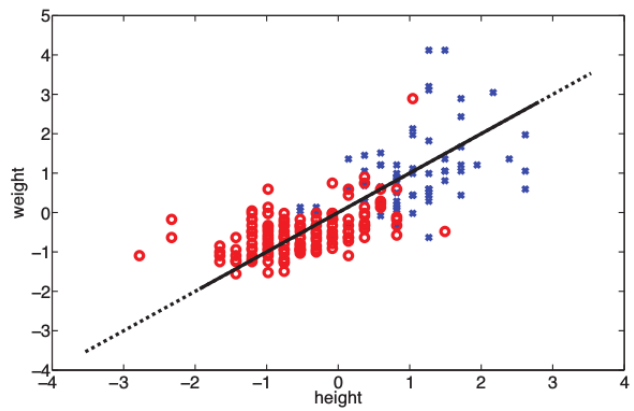
PCA

- the **principal directions** are the ones along which the data shows **maximal variance**
- this means that PCA can be "misled" by directions in which the variance is high merely because of the **measurement scale**
- in the figure below, the vertical axis (weight) uses a large range than the horizontal axis (height), resulting in a line that looks somewhat "unnatural"
- it is therefore standard practice to **standardize the data** first

$$x_{ij} \rightarrow \frac{(x_{ij} - \mu_j)}{\sigma_j}$$

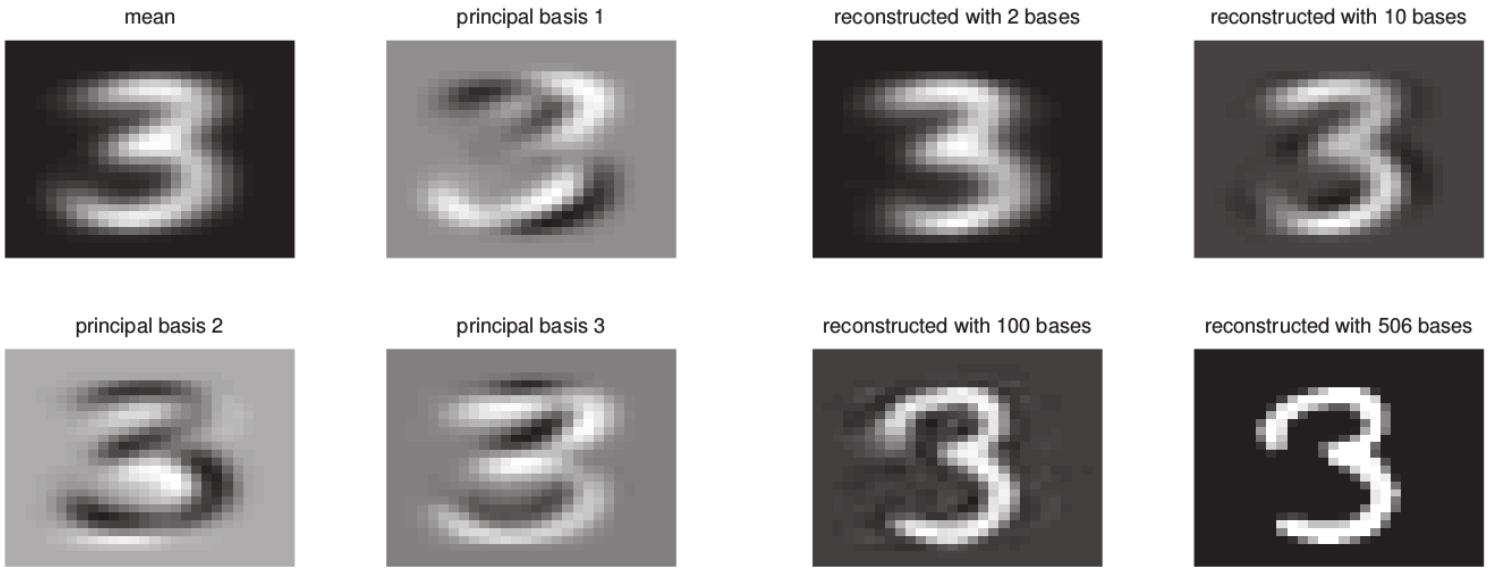
where $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ and $\sigma_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2$





Principal Component Analysis

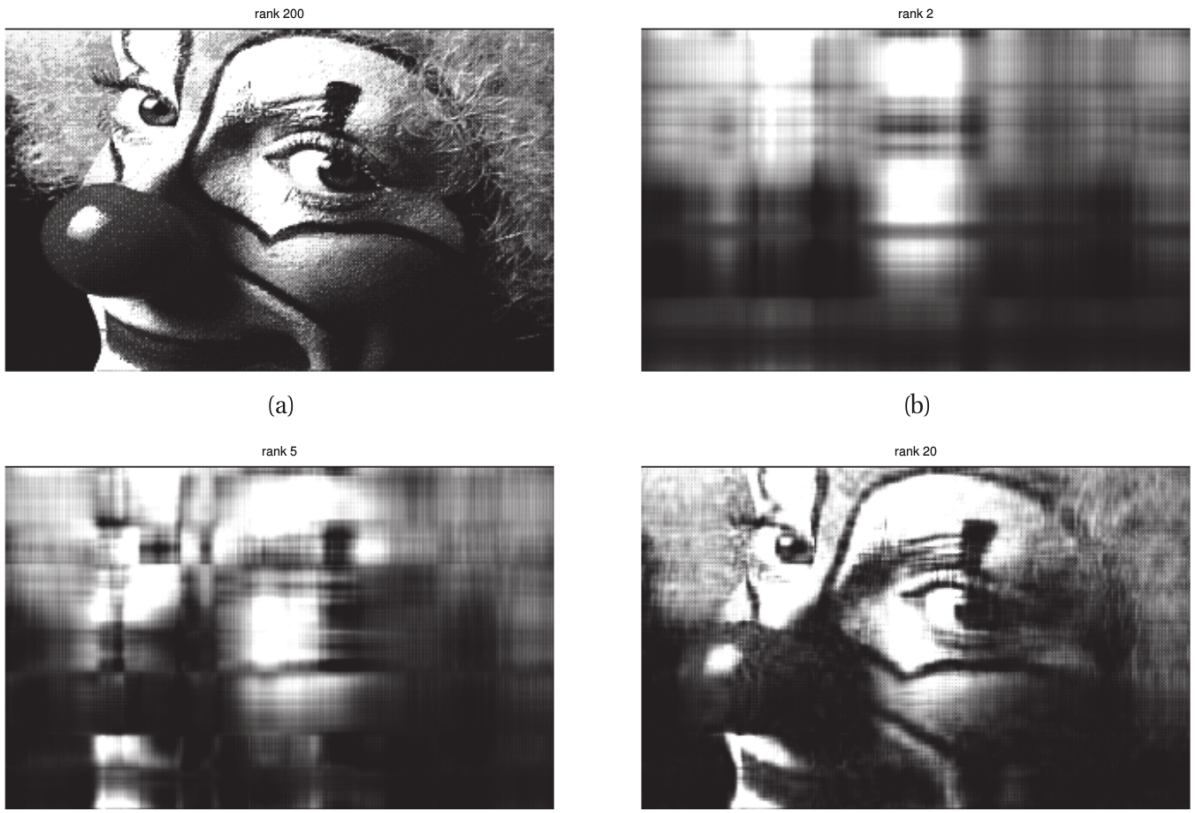
PCA



- **left**: the mean and the first three PC basis vectors (eigendigits) based on 25 images of the digit 3 (from the MNIST dataset)
- **right**: reconstruction of an image based on 2, 10, 100 and all the basis vectors

Principal Component Analysis

PCA



- low rank approximations to an image
- **top left**: the original image is of size 200×320, so has rank 200
- subsequent images have ranks 2, 5, and 20.

Credits

- Kevin Murphy's book