More on SVD and PCA

Rafikul Alam
Department of Mathematics
Indian Institute of Technology Guwahati
Guwahati - 781039, INDIA

Outline

- More on SVD
- Low rank approximation
- SVD and PCA

More on SVD

Let $A \in \mathbb{C}^{m\times n}$ and $A = U\Sigma V^*$ be an SVD. Then we have

$$\begin{aligned} AV &= U\Sigma, & A^*U &= V\Sigma \\ AV &= U\Sigma, & -A^*U &= -V\Sigma \end{aligned}$$

More on SVD

Let $A \in \mathbb{C}^{m\times n}$ and $A = U\Sigma V^*$ be an SVD. Then we have

$$\begin{matrix} AV = U\Sigma, & A^*U = V\Sigma \\ AV = U\Sigma, & -A^*U = -V\Sigma \end{matrix} \implies \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}.$$

More on SVD

Let $A \in \mathbb{C}^{m \times n}$ and $A = U\Sigma V^*$ be an SVD. Then we have

$$\begin{array}{ll} AV = U\Sigma, & A^*U = V\Sigma \\ AV = U\Sigma, & -A^*U = -V\Sigma \end{array} \Longrightarrow \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}.$$

Define

$$\mathbb{A} := \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)} \text{ and } \mathbb{U} := \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)}.$$

More on SVD

Let $A \in \mathbb{C}^{m\times n}$ and $A = U\Sigma V^*$ be an SVD. Then we have

$$\begin{matrix} AV = U\Sigma, & A^*U = V\Sigma \\ AV = U\Sigma, & -A^*U = -V\Sigma \end{matrix} \implies \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}.$$

Define

$$\mathbb{A} := \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)} \text{ and } \mathbb{U} := \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)}.$$

Then $\mathbb{A}$ is Hermitian and $\mathbb{U}$ is unitary. Further, we have the spectral decomposition

$$\mathbb{A} = \mathbb{U} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix} \mathbb{U}^*.$$

More on SVD

Let $A \in \mathbb{C}^{m\times n}$ and $A = U\Sigma V^*$ be an SVD. Then we have

$$\begin{aligned} AV &= U\Sigma, & A^*U &= V\Sigma \\ AV &= U\Sigma, & -A^*U &= -V\Sigma \end{aligned} \implies \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}.$$

Define

$$\mathbb{A} := \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)} \text{ and } \mathbb{U} := \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)}.$$

Then $\mathbb{A}$ is Hermitian and $\mathbb{U}$ is unitary. Further, we have the spectral decomposition

$$\mathbb{A} = \mathbb{U} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix} \mathbb{U}^*.$$

Let $V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$ and $U = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}$. Suppose that $\mathrm{rank}(A) = r$. Then

More on SVD

Let $A \in \mathbb{C}^{m\times n}$ and $A = U\Sigma V^*$ be an SVD. Then we have

$$\begin{matrix} AV = U\Sigma, & A^*U = V\Sigma \\ AV = U\Sigma, & -A^*U = -V\Sigma \end{matrix} \implies \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}.$$

Define

$$\mathbb{A} := \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)} \text{ and } \mathbb{U} := \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+n)}.$$

Then $\mathbb{A}$ is Hermitian and $\mathbb{U}$ is unitary. Further, we have the spectral decomposition

$$\mathbb{A} = \mathbb{U} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix} \mathbb{U}^*.$$

Let $V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$ and $U = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}$. Suppose that $\mathrm{rank}(A) = r$. Then

$$\mathbb{A} \begin{bmatrix} v_i \\ u_i \end{bmatrix} = \sigma_i \begin{bmatrix} v_i \\ u_i \end{bmatrix} \quad \text{and} \quad \mathbb{A} \begin{bmatrix} v_i \\ -u_i \end{bmatrix} = -\sigma_i \begin{bmatrix} v_i \\ -u_i \end{bmatrix} \quad \text{for } i = 1:r.$$

Extreme singular values

Let $A \in \mathbb{C}^{m\times n}$. Consider the SVD $A = U\Sigma V^*$. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, respectively, denote the largest and the smallest nonzero singular values of A. Then

$$\sigma_{\max}(A) = \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

is the maximum magnification of a vector by A.

Extreme singular values

Let $A \in \mathbb{C}^{m\times n}$. Consider the SVD $A = U\Sigma V^*$. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, respectively, denote the largest and the smallest nonzero singular values of A. Then

$$\sigma_{\max}(A) = \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

is the maximum magnification of a vector by A. Indeed, $\|A\|_2 = \|U\Sigma V^*\|_2 = \|\Sigma\|_2 = \sigma_{\max}(A)$.

Extreme singular values

Let $A \in \mathbb{C}^{m\times n}$. Consider the SVD $A = U\Sigma V^*$. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, respectively, denote the largest and the smallest nonzero singular values of A. Then

$$\sigma_{\max}(A) = \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

is the maximum magnification of a vector by A. Indeed, $\|A\|_2 = \|U\Sigma V^*\|_2 = \|\Sigma\|_2 = \sigma_{\max}(A)$.

Similarly,

$$\sigma_{\min}(A) = \min_{\|x\|_2=1} \{\|Ax\|_2 : x \notin N(A)\}$$

is the minimum nonzero magnification of a vector by A.

Extreme singular values

Let $A \in \mathbb{C}^{m\times n}$. Consider the SVD $A = U\Sigma V^*$. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, respectively, denote the largest and the smallest nonzero singular values of A. Then

$$\sigma_{\max}(A) = \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

is the maximum magnification of a vector by A. Indeed, $\|A\|_2 = \|U\Sigma V^*\|_2 = \|\Sigma\|_2 = \sigma_{\max}(A)$.

Similarly,

$$\sigma_{\min}(A) = \min_{\|x\|_2=1} \{\|Ax\|_2 : x \notin N(A)\}$$

is the minimum nonzero magnification of a vector by A.

Indeed, $\|Ax\|_2 = \|U\Sigma V^* x\|_2 = \|\Sigma y\|_2 \Longrightarrow \min_{\|y\|_2=1} \|\Sigma y\|_2 = \sigma_{\min}(A)$ for $y \notin N(\Sigma)$. Thus $\sigma_{\min}(A)\|x\|_2 \leq \|Ax\|_2 \leq \sigma_{\max}(A)\|x\|_2$ for $x \notin N(A)$.

Extreme singular values

Let $A \in \mathbb{C}^{m\times n}$. Consider the SVD $A = U\Sigma V^*$. Let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, respectively, denote the largest and the smallest nonzero singular values of A. Then

$$\sigma_{\max}(A) = \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

is the maximum magnification of a vector by A. Indeed, $\|A\|_2 = \|U\Sigma V^*\|_2 = \|\Sigma\|_2 = \sigma_{\max}(A)$.

Similarly,

$$\sigma_{\min}(A) = \min_{\|x\|_2=1} \{\|Ax\|_2 : x \notin N(A)\}$$

is the minimum nonzero magnification of a vector by A.

Indeed, $\|Ax\|_2 = \|U\Sigma V^* x\|_2 = \|\Sigma y\|_2 \Longrightarrow \min_{\|y\|_2=1} \|\Sigma y\|_2 = \sigma_{\min}(A)$ for $y \notin N(\Sigma)$. Thus $\sigma_{\min}(A)\|x\|_2 \leq \|Ax\|_2 \leq \sigma_{\max}(A)\|x\|_2$ for $x \notin N(A)$.

Recall that the condition number of A is given by

$$\mathrm{cond}_2(A) := \|A\|_2\|A^+\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}.$$

Properties of singular values

Theorem: Let $A \in \mathbb{C}^{m\times n}$. Let $\sigma_1(A) \geq \cdots \geq \sigma_p(A)$ be the singular values of A, where $p := \min(m, n)$. Then, for $k = 1 : p$, we have

$$\sigma_k(A) = \max_{\dim(S)=k} \min \left\{ \frac{\|Ax\|_2}{\|x\|_2} : x \in S \text{ and } x \neq 0 \right\}.$$

Let $E \in \mathbb{C}^{m\times n}$ and let $\sigma_1(A+E) \geq \cdots \geq \sigma_p(A+E)$ be the singular values of $A+E$. Then, for $k = 1 : p$, we have

$$|\sigma_k(A+E) - \sigma_k(A)| \leq \|E\|_2.$$

Singular values and singular vectors

Let $A \in \mathbb{C}^{m\times n}$ and $\mathrm{rank}(A) = r$. Then $A = U\Sigma V^* = \sum_{j=1}^{r} \sigma_j u_j v_j^*$, where

$$
\begin{array}{lll}
v_1 = \arg\max\limits_{\|x\|_2=1} \{\|Ax\|_2 : x \in \mathbb{R}^n\} & \sigma_1 := \|Av_1\|_2 & u_1 := Av_1/\sigma_1 \\
v_2 = \arg\max\limits_{\|x\|_2=1} \{\|Ax\|_2 : x \perp v_1\} & \sigma_2 := \|Av_2\|_2 & u_2 := Av_2/\sigma_2 \\
\vdots & \vdots & \vdots \\
v_r = \arg\max\limits_{\|x\|_2=1} \{\|Ax\|_2 : x \perp \{v_1, \ldots, v_{r-1}\}\} & \sigma_r := \|Av_r\|_2 & u_r := Av_r/\sigma_r
\end{array}
$$

Singular values and singular vectors

Let $A \in \mathbb{C}^{m\times n}$ and $\mathrm{rank}(A) = r$. Then $A = U\Sigma V^* = \sum_{j=1}^{r} \sigma_j u_j v_j^*$, where

$$
\begin{array}{lll}
v_1 = \arg\max\limits_{\|x\|_2=1} \{\|Ax\|_2 : x \in \mathbb{R}^n\} & \sigma_1 := \|Av_1\|_2 & u_1 := Av_1/\sigma_1 \\
v_2 = \arg\max\limits_{\|x\|_2=1} \{\|Ax\|_2 : x \perp v_1\} & \sigma_2 := \|Av_2\|_2 & u_2 := Av_2/\sigma_2 \\
\vdots & \vdots & \vdots \\
v_r = \arg\max\limits_{\|x\|_2=1} \{\|Ax\|_2 : x \perp \{v_1, \ldots, v_{r-1}\}\} & \sigma_r := \|Av_r\|_2 & u_r := Av_r/\sigma_r
\end{array}
$$

Also note that for $j = 1 : r$, we have

$$v_j = \arg\max_{\|x\|_2=1} \{\|Ax\|_2 : x \perp \{v_1, \ldots, v_{j-1}\}\} = \arg\max_{\|x\|_2=1} \{x^* A^* Ax : x \perp \{v_1, \ldots, v_{j-1}\}\}.$$

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$A_\ell \quad = \quad \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2,$$

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell}\|A - X\|_2, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell}\|A - X\|_F. \end{aligned}$$

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F. \end{aligned}$$

Theorem (Eckart-Young): Let $A \in \mathbb{C}^{m\times n}$ and $\ell < r := \mathrm{rank}(A)$.

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F. \end{aligned}$$

Theorem (Eckart-Young): Let $A \in \mathbb{C}^{m\times n}$ and $\ell < r := \mathrm{rank}(A)$. Let $A = \sum_{j=1}^{r} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix} V^*$ be an SVD of A.

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F. \end{aligned}$$

Theorem (Eckart-Young): Let $A \in \mathbb{C}^{m\times n}$ and $\ell < r := \mathrm{rank}(A)$. Let $A = \sum_{j=1}^{r} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix} V^*$ be an SVD of A.

Define $A_\ell := \sum_{j=1}^{\ell} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_\ell) & 0 \\ 0 & 0 \end{bmatrix} V^*$.

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F. \end{aligned}$$

Theorem (Eckart-Young): Let $A \in \mathbb{C}^{m\times n}$ and $\ell < r := \mathrm{rank}(A)$. Let $A = \sum_{j=1}^{r} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix} V^*$ be an SVD of A.

Define $A_\ell := \sum_{j=1}^{\ell} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_\ell) & 0 \\ 0 & 0 \end{bmatrix} V^*$. Then

$$A_\ell = \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \text{ and } \|A - A_\ell\|_2 = \sigma_{\ell+1},$$

Low rank approximation

Low rank approximation is used in many applications (e.g., data compression, pattern recognition, face detection, datamining and machine learning).

Task: Given $A \in \mathbb{C}^{m\times n}$ and $\ell < \mathrm{rank}(A)$, solve the minimation problems

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F. \end{aligned}$$

Theorem (Eckart-Young): Let $A \in \mathbb{C}^{m\times n}$ and $\ell < r := \mathrm{rank}(A)$. Let $A = \sum_{j=1}^{r} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix} V^*$ be an SVD of A.

Define $A_\ell := \sum_{j=1}^{\ell} \sigma_j u_j v_j^* = U \begin{bmatrix} \mathrm{diag}(\sigma_1, \cdots, \sigma_\ell) & 0 \\ 0 & 0 \end{bmatrix} V^*$. Then

$$\begin{aligned} A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \text{ and } \|A - A_\ell\|_2 = \sigma_{\ell+1}, \\ A_\ell &= \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F \text{ and } \|A - A_\ell\|_F = \sqrt{\sigma_{\ell+1}^2 + \cdots + \sigma_r^2}. \end{aligned}$$

Proof of Eckart-Young theorem

Obviously we have $\min_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Proof of Eckart-Young theorem

Obviously we have $\min_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Suppose that there exists X such that $\mathrm{rank}(X) = \ell$ and $\|A - X\|_2 < \sigma_{\ell+1}$. Then for any

Proof of Eckart-Young theorem

Obviously we have $\min_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Suppose that there exists X such that $\mathrm{rank}(X) = \ell$ and $\|A - X\|_2 < \sigma_{\ell+1}$. Then for any

$$\|u\|_2 = 1 \text{ and } u \in N(X) \Longrightarrow \|Au\|_2 = \|(A - X)u\|_2 < \sigma_{\ell+1}.$$

Proof of Eckart-Young theorem

Obviously we have $\min_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Suppose that there exists X such that $\mathrm{rank}(X) = \ell$ and $\|A - X\|_2 < \sigma_{\ell+1}$. Then for any

$$\|u\|_2 = 1 \text{ and } u \in N(X) \Longrightarrow \|Au\|_2 = \|(A - X)u\|_2 < \sigma_{\ell+1}.$$

Consider the subspace $S := \mathrm{span}(v_1, \cdots, v_{\ell+1})$. Then $S \cap N(X) \neq \{0\}$ (Why?). Hence there exists a nonzero $u \in S \cap N(X)$ such that $\|u\|_2 = 1$.

Proof of Eckart-Young theorem

Obviously we have $\min_{\operatorname{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Suppose that there exists X such that $\operatorname{rank}(X) = \ell$ and $\|A - X\|_2 < \sigma_{\ell+1}$. Then for any

$$\|u\|_2 = 1 \text{ and } u \in N(X) \Longrightarrow \|Au\|_2 = \|(A - X)u\|_2 < \sigma_{\ell+1}.$$

Consider the subspace $S := \operatorname{span}(v_1, \cdots, v_{\ell+1})$. Then $S \cap N(X) \neq \{0\}$ (Why?). Hence there exists a nonzero $u \in S \cap N(X)$ such that $\|u\|_2 = 1$.

Now $u \in S \Longrightarrow u = \alpha_1 v_1 + \cdots + \alpha_{\ell+1} v_{\ell+1} \Longrightarrow \|u\|_2 = \sqrt{|\alpha_1|^2 + \cdots + |\alpha_{\ell+1}|^2} = 1.$

Proof of Eckart-Young theorem

Obviously we have $\min_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Suppose that there exists X such that $\mathrm{rank}(X) = \ell$ and $\|A - X\|_2 < \sigma_{\ell+1}$. Then for any

$$\|u\|_2 = 1 \text{ and } u \in N(X) \implies \|Au\|_2 = \|(A - X)u\|_2 < \sigma_{\ell+1}.$$

Consider the subspace $S := \mathrm{span}(v_1, \cdots, v_{\ell+1})$. Then $S \cap N(X) \neq \{0\}$ (Why?). Hence there exists a nonzero $u \in S \cap N(X)$ such that $\|u\|_2 = 1$.

Now $u \in S \implies u = \alpha_1 v_1 + \cdots + \alpha_{\ell+1} v_{\ell+1} \implies \|u\|_2 = \sqrt{|\alpha_1|^2 + \cdots + |\alpha_{\ell+1}|^2} = 1$. This shows that

$$Au = \sigma_1 \alpha_1 u_1 + \cdots + \sigma_{\ell+1}\alpha_{\ell+1} u_{\ell+1} \implies \|Au\|_2 \geq \sigma_{\ell+1}\|u\|_2 \geq \sigma_{\ell+1}$$

which contradicts that $\|Au\|_2 < \sigma_{\ell+1}$ for any $u \in N(X)$ such that $\|u\|_2 = 1$. ■

Proof of Eckart-Young theorem

Obviously we have $\min_{\mathrm{rank}(X)=\ell} \|A - X\|_2 \leq \|A - A_\ell\|_2 = \sigma_{\ell+1}$.

Suppose that there exists X such that $\mathrm{rank}(X) = \ell$ and $\|A - X\|_2 < \sigma_{\ell+1}$. Then for any

$$\|u\|_2 = 1 \text{ and } u \in N(X) \Longrightarrow \|Au\|_2 = \|(A - X)u\|_2 < \sigma_{\ell+1}.$$

Consider the subspace $S := \mathrm{span}(v_1, \cdots, v_{\ell+1})$. Then $S \cap N(X) \neq \{0\}$ (Why?). Hence there exists a nonzero $u \in S \cap N(X)$ such that $\|u\|_2 = 1$.

Now $u \in S \Longrightarrow u = \alpha_1 v_1 + \cdots + \alpha_{\ell+1} v_{\ell+1} \Longrightarrow \|u\|_2 = \sqrt{|\alpha_1|^2 + \cdots + |\alpha_{\ell+1}|^2} = 1$. This shows that

$$Au = \sigma_1 \alpha_1 u_1 + \cdots + \sigma_{\ell+1} \alpha_{\ell+1} u_{\ell+1} \Longrightarrow \|Au\|_2 \geq \sigma_{\ell+1} \|u\|_2 \geq \sigma_{\ell+1}$$

which contradicts that $\|Au\|_2 < \sigma_{\ell+1}$ for any $u \in N(X)$ such that $\|u\|_2 = 1$. ■

Remark: The proof for the Frobenius norm follows from the fact that

$$\|A - A_\ell\|_F = \sqrt{\sigma_{\ell+1}^2 + \cdots + \sigma_r^2} \text{ and } \|Au\|_2 \leq \|A\|_F \|u\|_2.$$

Further, $A_\ell = \mathrm{argmin}_{\mathrm{rank}(X)=\ell} \|A - X\|_F$ is unique.

Consequences of Eckart-Young theorem

Suppose $A \in \mathbb{C}^{n\times n}$ is nonsingular. Consider the SVD $A = U\mathrm{diag}(\sigma_1, \cdots, \sigma_n)V^*$. Set $A_{n-1} := U\mathrm{diag}(\sigma_1, \cdots, \sigma_{n-1}, 0)V^*$. Then A_{n-1} is singular and

$$\sigma_n = \min\{\|A - X\|_2 : X \in \mathbb{C}^{n\times n},\ \mathrm{rank}(X) = n-1\} = \|A - A_{n-1}\|_2.$$

Consequences of Eckart-Young theorem

Suppose $A \in \mathbb{C}^{n\times n}$ is nonsingular. Consider the SVD $A = U\mathrm{diag}(\sigma_1, \cdots, \sigma_n)V^*$. Set $A_{n-1} := U\mathrm{diag}(\sigma_1, \cdots, \sigma_{n-1}, 0)V^*$. Then A_{n-1} is singular and

$$\sigma_n = \min\{\|A - X\|_2 : X \in \mathbb{C}^{n\times n},\ \mathrm{rank}(X) = n-1\} = \|A - A_{n-1}\|_2.$$

This shows that σ_n is the distance from A to the nearest singular matrix and that A_{n-1} is a nearest singular matrix. Hence σ_n is a measure of how close A to being a singular matrix.

Consequences of Eckart-Young theorem

Suppose $A \in \mathbb{C}^{n\times n}$ is nonsingular. Consider the SVD $A = U\mathrm{diag}(\sigma_1, \cdots, \sigma_n)V^*$. Set $A_{n-1} := U\mathrm{diag}(\sigma_1, \cdots, \sigma_{n-1}, 0)V^*$. Then A_{n-1} is singular and

$$\sigma_n = \min\{\|A - X\|_2 : X \in \mathbb{C}^{n\times n},\ \mathrm{rank}(X) = n-1\} = \|A - A_{n-1}\|_2.$$

This shows that σ_n is the distance from A to the nearest singular matrix and that A_{n-1} is a nearest singular matrix. Hence σ_n is a measure of how close A to being a singular matrix.

Remark: Note that $\det(A)$ is NOT a good measure of how close A to being singular. For example, if $A := \mathrm{diag}(1/2, \cdots, 1/2)$ then $\det(A) = 1/2^n$ but $\sigma_n = 1/2$.

Consequences of Eckart-Young theorem

Suppose $A \in \mathbb{C}^{n\times n}$ is nonsingular. Consider the SVD $A = U\mathrm{diag}(\sigma_1, \cdots, \sigma_n)V^*$. Set $A_{n-1} := U\mathrm{diag}(\sigma_1, \cdots, \sigma_{n-1}, 0)V^*$. Then A_{n-1} is singular and

$$\sigma_n = \min\{\|A - X\|_2 : X \in \mathbb{C}^{n\times n}, \ \mathrm{rank}(X) = n-1\} = \|A - A_{n-1}\|_2.$$

This shows that σ_n is the distance from A to the nearest singular matrix and that A_{n-1} is a nearest singular matrix. Hence σ_n is a measure of how close A to being a singular matrix.

Remark: Note that $\det(A)$ is NOT a good measure of how close A to being singular. For example, if $A := \mathrm{diag}(1/2, \cdots, 1/2)$ then $\det(A) = 1/2^n$ but $\sigma_n = 1/2$. Next, consider

$$A := \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 1 & -1 & \cdots & -1 & -1 \\ 0 & 0 & 1 & \cdots & -1 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, B := \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 1 & -1 & \cdots & -1 & -1 \\ 0 & 0 & 1 & \cdots & -1 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ \frac{-1}{2^{n-2}} & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

Consequences of Eckart-Young theorem

Then $\det(A) = 1$ and $Bx = 0$, where $x := \begin{bmatrix} 2^{n-2} & 2^{n-3} & \cdots & 2^0 & 1 \end{bmatrix}^\top$. Hence

$$\det(B) = 0 \text{ and } \sigma_n = \min\{\|A - X\|_2 : X \text{ singular}\} \leq \|A - B\|_2 = \frac{1}{2^{n-2}}.$$

This shows that A is close to being singular when n is large even though $\det(A) = 1$.

Consequences of Eckart-Young theorem

Then $\det(A) = 1$ and $Bx = 0$, where $x := \begin{bmatrix} 2^{n-2} & 2^{n-3} & \cdots & 2^0 & 1 \end{bmatrix}^\top$. Hence

$$\det(B) = 0 \text{ and } \sigma_n = \min\{\|A - X\|_2 : X \text{ singular}\} \leq \|A - B\|_2 = \frac{1}{2^{n-2}}.$$

This shows that A is close to being singular when n is large even though $\det(A) = 1$.

Numerical rank: If $A \in \mathbb{C}^{m \times n}$ is close enough to a matrix of rank r, where $r < \min(m, n)$, then A will behave like a rank r matrix in finite precision arithmetic. More precisely, the numerical rank of A is the number of singular values of A that are greater than $\max(m, n)\sigma_1 \mathbf{eps}$.

Consequences of Eckart-Young theorem

Then $\det(A) = 1$ and $Bx = 0$, where $x := \begin{bmatrix} 2^{n-2} & 2^{n-3} & \cdots & 2^0 & 1 \end{bmatrix}^\top$. Hence

$$\det(B) = 0 \text{ and } \sigma_n = \min\{\|A - X\|_2 : X \text{ singular}\} \leq \|A - B\|_2 = \frac{1}{2^{n-2}}.$$

This shows that A is close to being singular when n is large even though $\det(A) = 1$.

Numerical rank: If $A \in \mathbb{C}^{m \times n}$ is close enough to a matrix of rank r, where $r < \min(m, n)$, then A will behave like a rank r matrix in finite precision arithmetic. More precisely, the numerical rank of A is the number of singular values of A that are greater than $\max(m, n)\sigma_1$**eps**.

Example: Suppose that A is a 5×5 matrix with singular values

$$\sigma_1 = 4, \sigma_2 = 1, \sigma_3 = 10^{-12}, \sigma_4 = 3.1 \times 10^{-14}, \sigma_5 = 2.6 \times 10^{-15}.$$

Assume that **eps** $= 5 \times 10^{-15}$. Then $\sigma_1 \max(m, n)$**eps** $= 4 \times 5 \times 5 \times 10^{-15} = 10^{-13}$. Since three singular values of A are greater than 10^{-13}, the numerical rank of A is 3. ■

Variance and covariance

Let $x \in \mathbb{R}^n$. Then $\mathbf{x} := x - \mathrm{mean}(x)\mathbf{e}$ is the vector of deviations from $\mathrm{mean}(x) := \dfrac{1}{n}\sum_{j=1}^{n} x_j$,

where $\mathbf{e} := [1, \cdots, 1]^\top$. The variance σ^2 is defined by

$$\sigma^2 := \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \mathrm{mean}(\mathrm{x}))^2 = \frac{\mathbf{x}^\top \mathbf{x}}{n-1}.$$

The standard deviation is given by σ.

Variance and covariance

Let $x \in \mathbb{R}^n$. Then $\mathbf{x} := x - \mathrm{mean}(x)\mathbf{e}$ is the vector of deviations from $\mathrm{mean}(x) := \frac{1}{n}\sum_{j=1}^{n} x_j$, where $\mathbf{e} := [1, \cdots, 1]^\top$. The variance σ^2 is defined by

$$\sigma^2 := \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \mathrm{mean}(\mathrm{x}))^2 = \frac{\mathbf{x}^\top \mathbf{x}}{n-1}.$$

The standard deviation is given by σ. The covariance of zero mean vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by

$$\mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) := \frac{\mathbf{x}_1^\top \mathbf{x}_2}{n-1}.$$

Variance and covariance

Let $x \in \mathbb{R}^n$. Then $\mathbf{x} := x - \mathrm{mean}(x)\mathbf{e}$ is the vector of deviations from $\mathrm{mean}(x) := \frac{1}{n}\sum_{j=1}^{n} x_j$, where $\mathbf{e} := [1, \cdots, 1]^\top$. The variance σ^2 is defined by

$$\sigma^2 := \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \mathrm{mean}(\mathrm{x}))^2 = \frac{\mathbf{x}^\top \mathbf{x}}{n-1}.$$

The standard deviation is given by σ. The covariance of zero mean vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by

$$\mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) := \frac{\mathbf{x}_1^\top \mathbf{x}_2}{n-1}.$$

More generally, let $X = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ be such that $\mathrm{mean}(\mathbf{x}_j) = 0$ for $j = 1 : n$. Then

$$S = \frac{1}{n-1} X^\top X = \frac{1}{n-1}\begin{bmatrix} \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_n) \\ \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_n) & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_n) & \cdots & \mathrm{cov}(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

is called the covariance matrix.

Principal component analysis (PCA)

A Statistical view of PCA: Let $\mathbf{x} \in \mathbb{R}^m$ a zero-mean multivariate random variable and $\mathbf{u} \in \mathbb{R}^m$. Then the variance of $\mathbf{u}^\top \mathbf{x} \in \mathbb{R}$ is given by

$$\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) = \mathbb{E}((\mathbf{u}^\top \mathbf{x})^2) = \mathbb{E}(\mathbf{u}^\top \mathbf{x}\mathbf{x}^\top \mathbf{u}) = \mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u},$$

where $\mathbb{E}(\mathbf{x}\mathbf{x}^\top) \in \mathbb{R}^{m\times m}$ is the covariance matrix.

Principal component analysis (PCA)

A Statistical view of PCA: Let $\mathbf{x} \in \mathbb{R}^m$ a zero-mean multivariate random variable and $\mathbf{u} \in \mathbb{R}^m$. Then the variance of $\mathbf{u}^\top \mathbf{x} \in \mathbb{R}$ is given by

$$\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) = \mathbb{E}((\mathbf{u}^\top \mathbf{x})^2) = \mathbb{E}(\mathbf{u}^\top \mathbf{x}\mathbf{x}^\top \mathbf{u}) = \mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u},$$

where $\mathbb{E}(\mathbf{x}\mathbf{x}^\top) \in \mathbb{R}^{m\times m}$ is the covariance matrix.

Given a natural number $n < m$, the n principal components $\mathbf{y} := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top \in \mathbb{R}^n$ of $\mathbf{x}$ are defined as n uncorrelated linear components of $\mathbf{x}$,

$$y_i := \mathbf{u}_i^\top \mathbf{x} \in \mathbb{R}, \ \mathbf{u}_i \in \mathbb{R}^m, \ \mathbf{u}_i^\top \mathbf{u}_i = 1, \ i = 1:n,$$

such that the variances of $y_1, \ldots, y_n$ are maximized and satisfy

Principal component analysis (PCA)

A Statistical view of PCA: Let $\mathbf{x} \in \mathbb{R}^m$ a zero-mean multivariate random variable and $\mathbf{u} \in \mathbb{R}^m$. Then the variance of $\mathbf{u}^\top \mathbf{x} \in \mathbb{R}$ is given by

$$\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) = \mathbb{E}((\mathbf{u}^\top \mathbf{x})^2) = \mathbb{E}(\mathbf{u}^\top \mathbf{x}\mathbf{x}^\top \mathbf{u}) = \mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u},$$

where $\mathbb{E}(\mathbf{x}\mathbf{x}^\top) \in \mathbb{R}^{m \times m}$ is the covariance matrix.

Given a natural number $n < m$, the n principal components $\mathbf{y} := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top \in \mathbb{R}^n$ of $\mathbf{x}$ are defined as n uncorrelated linear components of $\mathbf{x}$,

$$y_i := \mathbf{u}_i^\top \mathbf{x} \in \mathbb{R}, \ \mathbf{u}_i \in \mathbb{R}^m, \ \mathbf{u}_i^\top \mathbf{u}_i = 1, \ i = 1 : n,$$

such that the variances of $y_1, \ldots, y_n$ are maximized and satisfy

$$\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_n) > 0.$$

The vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are called principal component directions.

A Statistical view of PCA

For example, the first principal component y_1 seeks to determine $\mathbf{u}_1$ such that

$$\begin{aligned}\mathbf{u}_1 &= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\} \\ &= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\}.\end{aligned}$$

A Statistical view of PCA

For example, the first principal component y_1 seeks to determine $\mathbf{u}_1$ such that

$$\begin{aligned}\mathbf{u}_1 &= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\} \\ &= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\}.\end{aligned}$$

The second principal component y_1 seeks to determine $\mathbf{u}_2$ such that

$$\begin{aligned}\mathbf{u}_2 &:= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m, \ \mathbf{u} \perp \mathbf{u}_1, \ \mathbf{u}^\top \mathbf{u} = 1\} \\ &= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m, \ \mathbf{u} \perp \mathbf{u}_1, \ \mathbf{u}^\top \mathbf{u} = 1\}.\end{aligned}$$

A Statistical view of PCA

For example, the first principal component y_1 seeks to determine $\mathbf{u}_1$ such that

$$\begin{aligned}
\mathbf{u}_1 &= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\} \\
&= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\}.
\end{aligned}$$

The second principal component y_1 seeks to determine $\mathbf{u}_2$ such that

$$\begin{aligned}
\mathbf{u}_2 &:= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m,\ \mathbf{u} \perp \mathbf{u}_1,\ \mathbf{u}^\top \mathbf{u} = 1\} \\
&= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m,\ \mathbf{u} \perp \mathbf{u}_1,\ \mathbf{u}^\top \mathbf{u} = 1\}.
\end{aligned}$$

Theorem: Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of a zero-mean multivariate random variable $\mathbf{x} \in \mathbb{R}^n$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$.

A Statistical view of PCA

For example, the first principal component y_1 seeks to determine $\mathbf{u}_1$ such that

$$\begin{aligned} \mathbf{u}_1 &= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\} \\ &= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m, \mathbf{u}^\top \mathbf{u} = 1\}. \end{aligned}$$

The second principal component y_1 seeks to determine $\mathbf{u}_2$ such that

$$\begin{aligned} \mathbf{u}_2 &:= \arg\max\{\mathrm{Var}(\mathbf{u}^\top \mathbf{x}) : \mathbf{u} \in \mathbb{R}^m,\ \mathbf{u} \perp \mathbf{u}_1,\ \mathbf{u}^\top \mathbf{u} = 1\} \\ &= \arg\max\{\mathbf{u}^\top \mathbb{E}(\mathbf{x}\mathbf{x}^\top)\mathbf{u} : \mathbf{u} \in \mathbb{R}^m,\ \mathbf{u} \perp \mathbf{u}_1,\ \mathbf{u}^\top \mathbf{u} = 1\}. \end{aligned}$$

Theorem: Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of a zero-mean multivariate random variable $\mathbf{x} \in \mathbb{R}^n$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Moreover, $\sigma_j^2 = \mathrm{Var}(y_j)$ for $j = 1 : n$. ■

A Statistical view of PCA and SVD

Sample Principal Components: The covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ of a zero-mean multivariate random variable $\mathbf{x} \in \mathbb{R}^m$ may not be known in practice. Instead, we may be given N i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of the random variable $\mathbf{x}$.

A Statistical view of PCA and SVD

Sample Principal Components: The covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ of a zero-mean multivariate random variable $\mathbf{x} \in \mathbb{R}^m$ may not be known in practice. Instead, we may be given N i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of the random variable $\mathbf{x}$.

Consider the data matrix $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m\times N}$. Note that each row of $\mathtt{X}$ has zero mean. The maximum likelihood estimate of $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ yields the sample covariance matrix $S_{\mathbf{x}}$, where

$$S_{\mathbf{x}} := \frac{1}{N-1}\sum_{j=1}^{N} \mathbf{x}_j \mathbf{x}_j^\top = \frac{1}{N-1}\mathtt{X}\mathtt{X}^\top.$$

A Statistical view of PCA and SVD

Sample Principal Components: The covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ of a zero-mean multivariate random variable $\mathbf{x} \in \mathbb{R}^m$ may not be known in practice. Instead, we may be given N i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of the random variable $\mathbf{x}$.

Consider the data matrix $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$. Note that each row of $\mathtt{X}$ has zero mean. The maximum likelihood estimate of $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ yields the sample covariance matrix $S_\mathbf{x}$, where

$$S_\mathbf{x} := \frac{1}{N-1} \sum_{j=1}^{N} \mathbf{x}_j \mathbf{x}_j^\top = \frac{1}{N-1} \mathtt{X}\mathtt{X}^\top.$$

The first n sample principal components $y_1, \ldots, y_n$ of the random variable $\mathbf{x}$ are defined as

$$y_i := \mathbf{u}_i^\top \mathbf{x} \text{ for } i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of $S_\mathbf{x} := \frac{1}{N-1}\mathtt{X}\mathtt{X}^\top$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2$.

A Statistical view of PCA and SVD

Sample Principal Components: The covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ of a zero-mean multivariate random variable $\mathbf{x} \in \mathbb{R}^m$ may not be known in practice. Instead, we may be given N i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of the random variable $\mathbf{x}$.

Consider the data matrix $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$. Note that each row of $\mathtt{X}$ has zero mean. The maximum likelihood estimate of $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ yields the sample covariance matrix $S_\mathbf{x}$, where

$$S_\mathbf{x} := \frac{1}{N-1} \sum_{j=1}^{N} \mathbf{x}_j \mathbf{x}_j^\top = \frac{1}{N-1} \mathtt{X}\mathtt{X}^\top.$$

The first n sample principal components $y_1, \ldots, y_n$ of the random variable $\mathbf{x}$ are defined as

$$y_i := \mathbf{u}_i^\top \mathbf{x} \text{ for } i = 1 : n,$$

As if X/sqrt(N-1) = USV*, then XX(T)/(N-1) = U[S^2, 0; 0, 0;]U*,
So XX(T)ui = sigma^2*ui

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of $S_\mathbf{x} := \frac{1}{N-1}\mathtt{X}\mathtt{X}^\top$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2$. Equivalently, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are left singular vectors of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$.

PCA and SVD

The left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$ are called sample principal component directions.

PCA and SVD

The left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$ are called sample principal component directions.We can consider SVD of $\mathtt{X}$ instead of $\mathtt{X}/\sqrt{N-1}$ and scale the singular values of $\mathtt{X}$.

PCA and SVD

The left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$ are called sample principal component directions.We can consider SVD of $\mathtt{X}$ instead of $\mathtt{X}/\sqrt{N-1}$ and scale the singular values of $\mathtt{X}$. Indeed, if $\sigma_1(\mathtt{X}), \ldots, \sigma_n(\mathtt{X})$ and $\sigma_1, \ldots, \sigma_n$ are the first n singular values of $\mathtt{X}$ and $\mathtt{X}/\sqrt{N-1}$, respectively, then

$$\sigma_j = \sigma_j(\mathtt{X})/\sqrt{N-1} \ \text{ for } \ j = 1:n.$$

PCA and SVD

The left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$ are called sample principal component directions.We can consider SVD of $\mathtt{X}$ instead of $\mathtt{X}/\sqrt{N-1}$ and scale the singular values of $\mathtt{X}$. Indeed, if $\sigma_1(\mathtt{X}), \ldots, \sigma_n(\mathtt{X})$ and $\sigma_1, \ldots, \sigma_n$ are the first n singular values of $\mathtt{X}$ and $\mathtt{X}/\sqrt{N-1}$, respectively, then

$$\sigma_j = \sigma_j(\mathtt{X})/\sqrt{N-1} \ \text{ for } \ j = 1 : n.$$

Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be left and right singular vectors of $\mathtt{X}$ corresponding to the first n singular values $\sigma_1(\mathtt{X}), \ldots, \sigma_1(\mathtt{X})$. Then $\mathtt{X}\mathbf{v}_i = \sigma_i(\mathtt{X})\mathbf{u}_i$ and $\mathbf{u}_i^\top \mathtt{X} = \sigma_i(\mathtt{X})\mathbf{v}_i^\top$ for $i = 1 : n$.

PCA and SVD

The left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$ are called sample principal component directions.We can consider SVD of $\mathtt{X}$ instead of $\mathtt{X}/\sqrt{N-1}$ and scale the singular values of $\mathtt{X}$. Indeed, if $\sigma_1(\mathtt{X}), \ldots, \sigma_n(\mathtt{X})$ and $\sigma_1, \ldots, \sigma_n$ are the first n singular values of $\mathtt{X}$ and $\mathtt{X}/\sqrt{N-1}$, respectively, then

$$\sigma_j = \sigma_j(\mathtt{X})/\sqrt{N-1} \text{ for } j = 1 : n.$$

Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be left and right singular vectors of $\mathtt{X}$ corresponding to the first n singular values $\sigma_1(\mathtt{X}), \ldots, \sigma_1(\mathtt{X})$. Then $\mathtt{X}\mathbf{v}_i = \sigma_i(\mathtt{X})\mathbf{u}_i$ and $\mathbf{u}_i^\top \mathtt{X} = \sigma_i(\mathtt{X})\mathbf{v}_i^\top$ for $i = 1 : n$. The row vector

$$\mathbf{y}_i := \mathbf{u}_i^\top \mathtt{X} = \sigma_i(\mathtt{X})\mathbf{v}_i^\top \in \mathbb{R}^{1\times N}$$

is the sample data of the principal component $y_i := \mathbf{u}_i^\top \mathbf{x}$ for $i = 1 : n$.

PCA and SVD

The left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathtt{X}/\sqrt{N-1}$ corresponding to the first n singular values $\sigma_1 \geq \cdots \geq \sigma_n$ are called sample principal component directions.We can consider SVD of $\mathtt{X}$ instead of $\mathtt{X}/\sqrt{N-1}$ and scale the singular values of $\mathtt{X}$. Indeed, if $\sigma_1(\mathtt{X}), \ldots, \sigma_n(\mathtt{X})$ and $\sigma_1, \ldots, \sigma_n$ are the first n singular values of $\mathtt{X}$ and $\mathtt{X}/\sqrt{N-1}$, respectively, then

$$\sigma_j = \sigma_j(\mathtt{X})/\sqrt{N-1} \;\text{ for }\; j = 1:n.$$

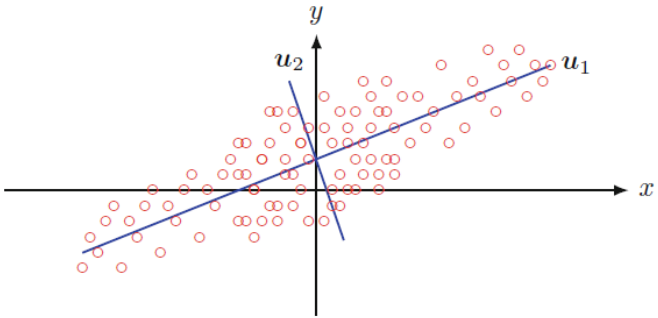
Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be left and right singular vectors of $\mathtt{X}$ corresponding to the first n singular values $\sigma_1(\mathtt{X}), \ldots, \sigma_1(\mathtt{X})$. Then $\mathtt{X}\mathbf{v}_i = \sigma_i(\mathtt{X})\mathbf{u}_i$ and $\mathbf{u}_i^\top \mathtt{X} = \sigma_i(\mathtt{X})\mathbf{v}_i^\top$ for $i = 1:n$. The row vector

$$\mathbf{y}_i := \mathbf{u}_i^\top \mathtt{X} = \sigma_i(\mathtt{X})\mathbf{v}_i^\top \in \mathbb{R}^{1\times N}$$

is the sample data of the principal component $y_i := \mathbf{u}_i^\top \mathbf{x}$ for $i = 1:n$. Moreover, for $i = 1:n$,

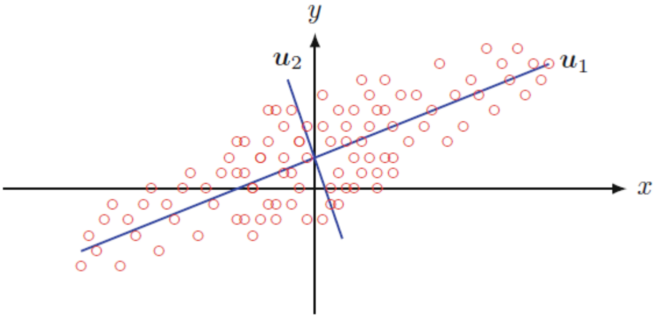
$$\mathrm{Var}(\mathbf{y}_i) = \frac{1}{N-1}\mathbf{y}_i^\top \mathbf{y}_i = \frac{\sigma_i(X)^2}{N-1}\mathbf{v}_i^\top \mathbf{v}_i = \frac{\sigma_i(X)^2}{N-1} = \sigma_i^2.$$

A Statistical view of PCA



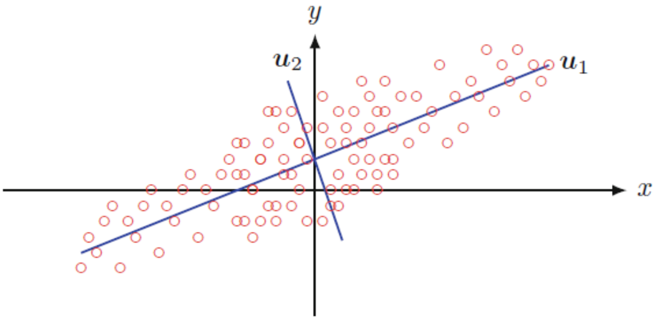
Remark: Computation of SVD of $\mathtt{X}$ is preferred over spectral decomposition of $\mathtt{X}\mathtt{X}^\top \in \mathbb{R}^{m\times m}$ due to finite precision arithmetic.

A Statistical view of PCA



Remark: Computation of SVD of $\mathtt{X}$ is preferred over spectral decomposition of $\mathtt{X}\mathtt{X}^\top \in \mathbb{R}^{m\times m}$ due to finite precision arithmetic. If $N < m$ then we can compute orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathtt{X}^\top\mathtt{X} \in \mathbb{R}^{N\times N}$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$ and set $\mathbf{u}_j := \mathtt{X}\mathbf{v}_j/\sigma_j$ for $j = 1 : n$.

A Statistical view of PCA



Remark: Computation of SVD of $\mathtt{X}$ is preferred over spectral decomposition of $\mathtt{X}\mathtt{X}^\top \in \mathbb{R}^{m\times m}$ due to finite precision arithmetic. If $N < m$ then we can compute orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathtt{X}^\top\mathtt{X} \in \mathbb{R}^{N\times N}$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$ and set $\mathbf{u}_j := \mathtt{X}\mathbf{v}_j/\sigma_j$ for $j = 1 : n$. For stable computation, it is advisable to compute SVD of $\mathtt{X}$.

PCA Summary

PCA of a Random Variable: Let $\mathbf{x} \in \mathbb{R}^m$ be a zero-mean multivariate random variable. Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$.

PCA Summary

PCA of a Random Variable: Let $\mathbf{x} \in \mathbb{R}^m$ be a zero-mean multivariate random variable. Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Moreover, $\sigma_j^2 = \mathrm{Var}(y_j)$ for $j = 1 : n$. ■

PCA Summary

PCA of a Random Variable: Let $\mathbf{x} \in \mathbb{R}^m$ be a zero-mean multivariate random variable. Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Moreover, $\sigma_j^2 = \mathrm{Var}(y_j)$ for $j = 1 : n$. ■

PCA of Samples: Let $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$ be a data matrix, where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are i.i.d. samples of the zero-mean random variable $\mathbf{x} \in \mathbb{R}^m$.

PCA Summary

PCA of a Random Variable: Let $\mathbf{x} \in \mathbb{R}^m$ be a zero-mean multivariate random variable. Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Moreover, $\sigma_j^2 = \mathrm{Var}(y_j)$ for $j = 1 : n$. ■

PCA of Samples: Let $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$ be a data matrix, where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are i.i.d. samples of the zero-mean random variable $\mathbf{x} \in \mathbb{R}^m$. Then the first n sample principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \text{ for } i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the sample covariance matrix $S_{\mathbf{x}} := \dfrac{\mathtt{X}\mathtt{X}^\top}{N-1}$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2$.

PCA Summary

PCA of a Random Variable: Let $\mathbf{x} \in \mathbb{R}^m$ be a zero-mean multivariate random variable. Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Moreover, $\sigma_j^2 = \mathrm{Var}(y_j)$ for $j = 1 : n$. ■

PCA of Samples: Let $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$ be a data matrix, where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are i.i.d. samples of the zero-mean random variable $\mathbf{x} \in \mathbb{R}^m$. Then the first n sample principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \text{ for } i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the sample covariance matrix $S_{\mathbf{x}} := \dfrac{\mathtt{X}\mathtt{X}^\top}{N-1}$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2$. Equivalently, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are left singular vectors of $\mathtt{X}$ corresponding to the first n singular values $\sigma_1(\mathtt{X}) \geq \cdots \geq \sigma_n(\mathtt{X})$.

PCA Summary

PCA of a Random Variable: Let $\mathbf{x} \in \mathbb{R}^m$ be a zero-mean multivariate random variable. Assume that $\mathrm{rank}(\mathbb{E}(\mathbf{x}\mathbf{x}^\top)) \geq n$. Then the first n principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \quad \text{for} \quad i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the covariance matrix $\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ corresponding to the n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Moreover, $\sigma_j^2 = \mathrm{Var}(y_j)$ for $j = 1 : n$. ■

PCA of Samples: Let $\mathtt{X} := \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$ be a data matrix, where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are i.i.d. samples of the zero-mean random variable $\mathbf{x} \in \mathbb{R}^m$. Then the first n sample principal components $y_1, \ldots, y_n$ of $\mathbf{x}$ are given by

$$y_i := \mathbf{u}_i^\top \mathbf{x} \text{ for } i = 1 : n,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are orthonormal eigenvectors of the sample covariance matrix $S_{\mathbf{x}} := \dfrac{\mathtt{X}\mathtt{X}^\top}{N-1}$ corresponding to n largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2$. Equivalently, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are left singular vectors of $\mathtt{X}$ corresponding to the first n singular values $\sigma_1(\mathtt{X}) \geq \cdots \geq \sigma_n(\mathtt{X})$. Moreover, we have $\sigma_i^2 = \sigma_i(\mathtt{X})^2/(N-1) = \mathrm{Var}(\mathbf{u}_i^\top \mathtt{X})$ for $i = 1 : n$. ■