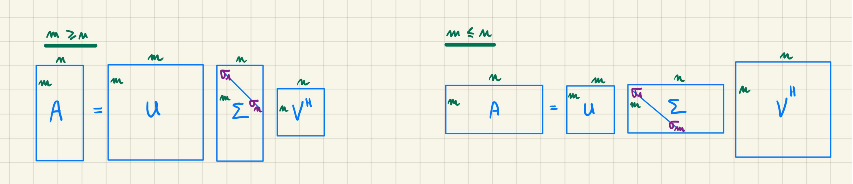
Singular Value Decomposition (SVD)

Any matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ can be factorized: $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{H}}$

- $\mathbf{U} \in \mathbb{C}^{m \times m}, \mathbf{V} \in \mathbb{C}^{n \times n}$ unitary matrices [1]
- $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_i)_{i=1}^{\min(m,n)}$ diagonal
- $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$ are **singular values**: unique and square roots of eigenvalues of $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ or $\mathbf{A}\mathbf{A}^{\mathrm{H}}$.



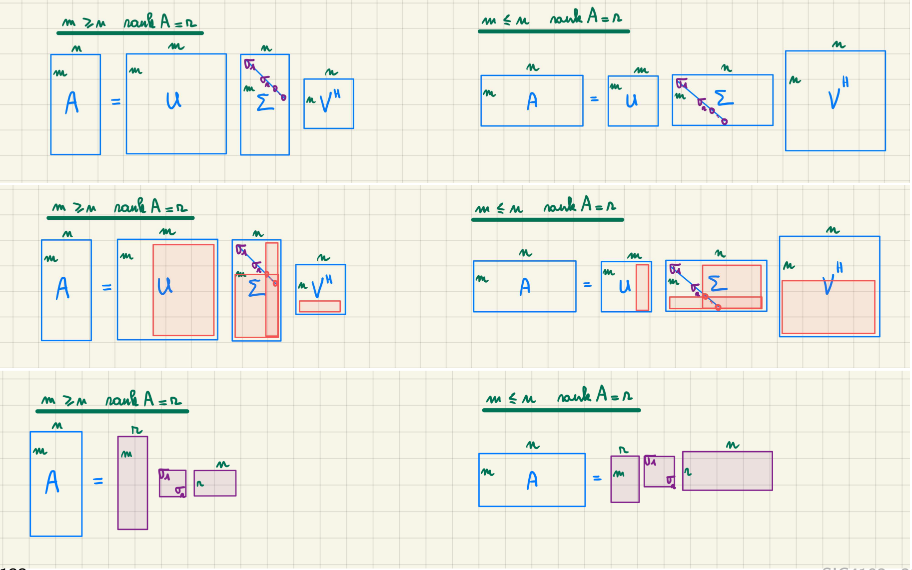
- With $\mathbf{U} = [\mathbf{u}_1 \ldots \mathbf{u}_m]$ and $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_n]$, sum of rank-1 matrices:

$$\mathbf{A} = \sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^{\mathrm{H}}$$

[1] $\mathbf{U}\mathbf{U}^{\mathrm{H}} = \mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{Id}_m$ and $\mathbf{V}\mathbf{V}^{\mathrm{H}} = \mathbf{V}^{\mathrm{H}}\mathbf{V} = \mathbf{Id}_n$

"Economy size" SVD

If $\operatorname{rank}\mathbf{A} = r$, then $\sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0$ and $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^{\mathrm{H}}$



Matrix norms

Write SVD decomposition: $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{H}}$ $\quad \boldsymbol{\Sigma} = \mathrm{Diag}(\boldsymbol{\sigma})$, with singular values vector $\boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_r \end{bmatrix}$.

- ℓ_2 (or **Schur/spectral**) norm: $\|\mathbf{A}\|_2 = \max_{i=1}^r \sigma_i$.

  Prop: $\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ is the **operator** norm.
- **Frobenius** norm: $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^r \sigma_i^2}$.

  Prop: $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}[\mathbf{A}^{\mathrm{H}}\mathbf{A}]}$ corresponds to scalar product $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}[\mathbf{X}^{\mathrm{H}}\mathbf{Y}]$.
- **nuclear** norm (or trace norm): $\|\mathbf{A}\|_* = \sum_{i=1}^r \sigma_i$

- Norm on matrix $\leftrightarrow$ norm on vector of singular values:
  $\|\mathbf{A}\|_2 = \|\boldsymbol{\sigma}\|_\infty \qquad \|\mathbf{A}\|_{\mathrm{F}} = \|\boldsymbol{\sigma}\|_2 \qquad \|\mathbf{A}\|_* = \|\boldsymbol{\sigma}\|_1$

Eckart-Young theorem

Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ or rank r and let $\|.\|$ be either $\|.\|_2$ or $\|.\|_{\mathrm{F}}$. Write $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{H}}$ the SVD.

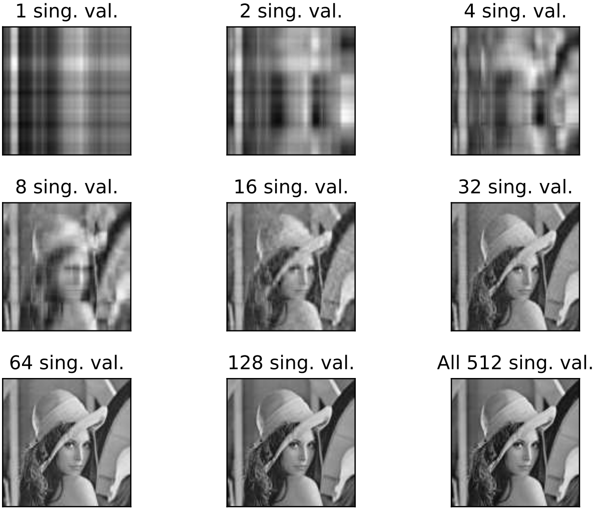
The solution to:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \|\mathbf{A} - \mathbf{X}\| \quad \text{s.t. } \operatorname{rank} \mathbf{X} \leq p$$

is given by $\mathbf{X}_p = \mathbf{U}\boldsymbol{\Sigma}_p\mathbf{V}^{\mathrm{H}}$ where $\boldsymbol{\Sigma}$ obtained from $\boldsymbol{\Sigma}$ by setting the $r - p$ smallest singular values to zero: $\sigma_{p+1} = \cdots = \sigma_r = 0$

Low-rank approximation

Example on on image



Maximizing variance

Random $\mathbf{x} \in \mathbb{R}^n$, centered, covariance $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{\Lambda} = \mathrm{Diag}(\lambda_i)_{i=1}^n$ and $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, $\mathbf{U}$ orthogonal.

Objective: find uncorrelated and maximal variance linear combinations of $\mathbf{x}$
$\leftrightsquigarrow$ find unit norm vectors $(\mathbf{w}_i)_{i=1}^p$ such that:

- $y_1 = \mathbf{w}_1{}^\top \mathbf{x}$: $\mathbb{E}\{y_1^2\}$ is maximal $\rightarrow \mathbf{w}_1 = \mathbf{u}_1$

Solution:
$\mathbb{E}\{y_1^2\} = \mathbf{w}_1{}^\top \mathbf{C}\mathbf{w}_1$ yields:

$$\mathbf{w}_1 = \arg\max_{|\mathbf{w}\|_2 = 1} \mathbf{w}_1{}^\top \mathbf{C}\mathbf{w}_1 = \mathbf{u}_1$$

Maximizing variance

Random $\mathbf{x} \in \mathbb{R}^n$, centered, covariance $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ with $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_i)_{i=1}^n$ and $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, $\mathbf{U}$ orthogonal.

Objective: find uncorrelated and maximal variance linear combinations of $\mathbf{x}$
$\leftrightsquigarrow$ find unit norm vectors $(\mathbf{w}_i)_{i=1}^p$ such that:

- $y_1 = \mathbf{w}_1{}^\top \mathbf{x}\ :\ \mathbb{E}\{y_1^2\}$ is maximal $\to \mathbf{w}_1 = \mathbf{u}_1$
- $y_2 = \mathbf{w}_2{}^\top \mathbf{x}\ :\ \mathbb{E}\{y_2 y_1\} = 0$ and $\mathbb{E}\{y_2^2\}$ is maximal $\to \mathbf{w}_2 = \mathbf{u}_2$

Solution:
$\mathbb{E}\{y_2^2\} = \mathbf{w}_2{}^\top \mathbf{C}\mathbf{w}_2$ and $\mathbb{E}\{y_2 y_1\} = \mathbf{w}_2{}^\top \mathbf{C}\mathbf{w}_1 = \lambda_1 \mathbf{w}_2{}^\top \mathbf{u}_1$ yield:

$$\mathbf{w}_2 = \arg \max_{|\mathbf{w}\|_2 = 1, \mathbf{w}^\top \mathbf{u}_1 = 0} \mathbf{w}_2{}^\top \mathbf{C}\mathbf{w}_2 = \mathbf{u}_2$$

Maximizing variance

Random $\mathbf{x} \in \mathbb{R}^n$, centered, covariance $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{\Lambda} = \mathrm{Diag}(\lambda_i)_{i=1}^n$ and $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, $\mathbf{U}$ orthogonal.

Objective: find uncorrelated and maximal variance linear combinations of $\mathbf{x}$
$\leftrightsquigarrow$ find unit norm vectors $(\mathbf{w}_i)_{i=1}^p$ such that:

- $y_1 = {\mathbf{w}_1}^\top \mathbf{x}$: $\mathbb{E}\{y_1^2\}$ is maximal $\to \mathbf{w}_1 = \mathbf{u}_1$
- $y_2 = {\mathbf{w}_2}^\top \mathbf{x}$: $\mathbb{E}\{y_2 y_1\} = 0$ and $\mathbb{E}\{y_2^2\}$ is maximal $\to \mathbf{w}_2 = \mathbf{u}_2$
- $y_3 = {\mathbf{w}_3}^\top \mathbf{x}$ such that: $\mathbb{E}\{y_1 y_3\} = \mathbb{E}\{y_2 y_3\} = 0$ and $\mathbb{E}\{y_3^2\}$ maximal $\to \mathbf{w}_3 = \mathbf{u}_3$
- ...

Solution:
$\mathbb{E}\{y_2^2\} = {\mathbf{w}_2}^\top \mathbf{C}\mathbf{w}_2$ and $\mathbb{E}\{y_2 y_1\} = {\mathbf{w}_2}^\top \mathbf{C}\mathbf{w}_1 = \lambda_1 {\mathbf{w}_2}^\top \mathbf{u}_1$ yield:

$$\mathbf{w}_2 = \arg \max_{|\mathbf{w}\|_2 = 1, \mathbf{w}^\top \mathbf{u}_1 = 0} {\mathbf{w}_2}^\top \mathbf{C}\mathbf{w}_2 = \mathbf{u}_2$$

Minimizing quadratic error

Random $\mathbf{x} \in \mathbb{R}^n$, centered, covariance $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\}$

Objective: find p-dimensional linear subspace $\subset \mathbb{R}^n$ such that projection of $\mathbf{x}$ minimizes quadratic error:

$$\min_{\mathbf{w}_1,\dots,\mathbf{w}_p} \mathbb{E}\left\{ \|\mathbf{x} - \sum_{i=1}^{p} (\mathbf{w}_i{}^\top \mathbf{x})\mathbf{w}_i\|_2^2 \right\}$$

where $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_p] \in \mathbb{R}^{n \times p}$ orthonormal basis ($\mathbf{W}^\top \mathbf{W} = \mathbf{Id}_p$).

Minimizing quadratic error

Random $\mathbf{x} \in \mathbb{R}^n$, centered, covariance $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\}$

Objective: find p-dimensional linear subspace $\subset \mathbb{R}^n$ such that projection of $\mathbf{x}$ minimizes quadratic error:

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_p} \mathbb{E}\left\{ \|\mathbf{x} - \sum_{i=1}^{p} (\mathbf{w}_i{}^\top \mathbf{x})\mathbf{w}_i\|_2^2 \right\}$$

where $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_p] \in \mathbb{R}^{n\times p}$ orthonormal basis ($\mathbf{W}^\top\mathbf{W} = \mathbf{Id}_p$).

$$\underbrace{\mathbb{E}\left\{ \|\mathbf{x} - \sum_{i=1}^{p} (\mathbf{w}_i{}^\top \mathbf{x})\mathbf{w}_i\|_2^2 \right\}}_{\text{minimize error}} = \mathrm{Tr}(\mathbf{C}) - \underbrace{\sum_{i=1}^{p} \mathbf{w}_i{}^\top \mathbf{C}\mathbf{w}_i}_{\text{maximize variance}}$$

➜ similar to previous problem, same solution.

Whitening

Random $\mathbf{x} \in \mathbb{R}^n$, centered, covariance $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{U} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \mathbf{U}^\top$

with $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, $\mathbf{U} = [\mathbf{u}_1 \ldots \mathbf{u}_n]$ orthogonal.

Let $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$.

- With PCA, $\mathbf{W} = [\mathbf{u}_1 \ldots \mathbf{u}_p]$:

$$\mathbb{E}\{\mathbf{y}\mathbf{y}^\top\} = \mathbf{W}^\top \mathbf{C} \mathbf{W} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$$

  Data has been decorrelated.

- With $\mathbf{W} = \left[\mathbf{u}_1 \ldots \mathbf{u}_p\right] \begin{bmatrix} \lambda_1^{-1/2} & & \\ & \ddots & \\ & & \lambda_p^{-1/2} \end{bmatrix}$:

$$\mathbb{E}\{\mathbf{y}\mathbf{y}^\top\} = \mathbf{Id}_p$$

  Data has been whitened.

Empirical data point of view

- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{n \times T}$: set of T vector samples
- Empirical covariance $\hat{\mathbf{C}} = \frac{1}{T}\mathbf{X}\mathbf{X}^\top$
- For any $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_p]$ with orthonormal columns:

$$\frac{1}{T}\sum_{t=1}^{T} \underbrace{\|\mathbf{x}_t - \sum_{i=1}^{p}(\mathbf{w}_i{}^\top \mathbf{x}_t)\mathbf{w}_i\|_2^2}_{\text{quadratic error}} = \mathrm{Tr}(\hat{\mathbf{C}}) - \frac{1}{T}\sum_{t=1}^{T} \underbrace{\|\mathbf{W}^\top \mathbf{x}_t\|_2^2}_{\text{norm of projection}}$$

➜ minimize quadratic error $\leftrightarrow$ maximize norm of projection

SVD based PCA

Compute "economy size" SVD of set of T vector samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{n \times T}$:

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

where $\operatorname{rank} \mathbf{X} = p$ and $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix}$

- Empirical covariance: $\hat{\mathbf{C}} = \frac{1}{T}\mathbf{X}\mathbf{X}^\top = \frac{1}{T}\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$

➜ PCA readily obtained (vectors in $\mathbf{U}$)

➜ $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}^{-1}$ is a whitening matrix and $\mathbf{Y} = \mathbf{W}^\top\mathbf{X}$

➜ If $p < n$, rows of $\mathbf{X}$ linearly dependent and $\mathbf{Y} \in \mathbb{R}^{p \times T}$: dimension reduction has been performed.

Example on MNIST dataset