Visualization
CH07: PCA

Jong-June Jeon

University of Seoul, Department of Statistics

Contents

- Singular Value Decomposition as an Optimal Low–Dimensional Approximation

- Principal Component Analysis

- Matrix and Linear Map

- Dimension Reduction and Optimal Reconstruction

- Appendix

Singular Value Decomposition as an Optimal Low–Dimensional Approximation

Singular Value Decomposition

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ ($p < n$) denote a (row-centered) data matrix whose rows are observations and columns are variables. The *singular value decomposition* (SVD) factorises $\mathbf{X}$ as

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

where

- $\mathbf{U} = \big[U_1, \ldots, U_p\big] \in \mathbb{R}^{n \times p}$ contains the left singular vectors $(\mathbf{U}^\top \mathbf{U} = \mathbf{I}_p)$;
- $\mathbf{V} = \big[V_1, \ldots, V_p\big] \in \mathbb{R}^{p \times p}$ contains the right singular vectors $(\mathbf{V}^\top \mathbf{V} = \mathbf{I}_p)$;
- $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$;
- $r = \mathrm{rank}(\mathbf{X}) \leq \min\{n, p\}$.

Singular Value Decomposition: Example I

Let

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

be a real matrix with rank $p = 2$. The **reduced singular value decomposition** (SVD) of $\mathbf{A}$ is given by

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

where

- $\mathbf{U} = \begin{bmatrix} -0.5026 & 0.7746 \\ -0.5740 & -0.6325 \\ -0.6464 & 0.0000 \end{bmatrix} \in \mathbb{R}^{3\times 2}$ contains the left singular vectors (orthonormal: $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_2$),
- $\mathbf{\Sigma} = \begin{bmatrix} 5.1962 & 0 \\ 0 & 1.7321 \end{bmatrix} \in \mathbb{R}^{2\times 2}$ is the diagonal matrix of singular values,

Singular Value Decomposition: Example II

- $\mathbf{V} = \begin{bmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \in \mathbb{R}^{2\times 2}$ contains the right singular vectors (orthonormal: $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_2$).

Thus, the matrix $\mathbf{A}$ can be approximately reconstructed as

$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \begin{bmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{bmatrix}.$$

Data and representation

$(A)_{ij}$ denotes the entry of A in the row i and the column j. For $i = 1, \cdots, n$,

$$x_i = (X)_{i1} e_1 + \cdots + (X)_{ip} e_p,$$

where $\{e_1, \ldots, e_p\}$ forms the standard basis of $\mathbb{R}^p$.

- Coordinate system: $(e_1, \cdots, e_p)$
- Scaling factor: $(1, \ldots, 1) \in \mathbb{R}^p$
- Representation of x_i w.r.t the $(e_1, \cdots, e_p)$: $((X)_{i1}, \cdots, (X)_{ip}) \in \mathbb{R}^p$.

How to obtain low dimensional representation of x_i effectively?

SVD: Definition & Structure

Full SVD (rank p)

$$\mathbf{X} = \underbrace{\mathbf{U}}_{n\times p}\,\underbrace{\mathbf{\Sigma}}_{p\times p}\,\underbrace{\mathbf{V}^\top}_{p\times p} = \sum_{i=1}^{p} \sigma_i\, U_i V_i^\top, \quad \sigma_1 \geq \cdots \geq \sigma_r > 0.$$

$$\mathbf{U\Sigma} = \begin{bmatrix} U_1 & \cdots & U_p \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix} = \begin{bmatrix} \sigma_1 U_1 & \cdots & \sigma U_p \end{bmatrix}$$

SVD: Definition & Structure

$$\mathbf{X} = \begin{bmatrix} \sigma_1 U_1 & \cdots & \sigma_p U_p \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \\ \vdots \\ V_p^\top \end{bmatrix} = \sigma_1 \; U_1 V_1^\top + \cdots + \sigma_p U_p V_p^\top$$

The example below helps understanding the above equation:

$$\underbrace{\begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ u_{13} & u_{23} \end{bmatrix}}_{\mathbf{U} \in \mathbb{R}^{3\times 2}} \underbrace{\begin{bmatrix} \sigma_1 v_{11} & \sigma_1 v_{21} \\ \sigma_2 v_{12} & \sigma_2 v_{22} \end{bmatrix}}_{\mathbf{\Sigma V}^\top \in \mathbb{R}^{2\times 2}} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}$$

$$\begin{aligned} &= \begin{bmatrix} \beta_{11} U_1 + \beta_{21} U_2 & \beta_{12} U_1 + \beta_{22} U_2 \end{bmatrix} \\ &= [\beta_{11} U_1 \;\; \beta_{12} U_1] + [\beta_{21} U_2 \;\; \beta_{22} U_2] \\ &= U_1 [\beta_{11} \;\; \beta_{12}] + U_2 [\beta_{21} \;\; \beta_{22}] \\ &= U_1 \sigma_1 V_1^\top + U_2 \sigma_2 V_2^\top = \sigma_1 U_1 V_1^\top + \sigma_2 U_2 V_2^\top . \end{aligned}$$

SVD: Definition & Structure

Our conclusion is that a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ has the following representation,

$$\mathbf{X} = \underbrace{\mathbf{U}}_{n \times p} \underbrace{\mathbf{\Sigma}}_{p \times p} \underbrace{\mathbf{V}^\top}_{p \times p} = \sum_{i=1}^{p} \sigma_i \, U_i V_i^\top, \quad \sigma_1 \geq \cdots \geq \sigma_r > 0.$$

SVD and representation

Denote the ith row vector of $\mathbf{X}$ by $x_i^\top$. Then,

$$\mathbf{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \begin{bmatrix} U_1 & \cdots & U_p \end{bmatrix} \underbrace{\begin{bmatrix} \sigma_1 V_1^\top \\ \cdots \\ \sigma_p V_p^\top \end{bmatrix}}_{=\mathbf{\Sigma}\mathbf{V}^\top}.$$

By taking transpose operator on $\mathbf{X}$

$$\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} = \begin{bmatrix} \sigma_1 V_1 & \cdots & \sigma_p V_p \end{bmatrix} \mathbf{U}^\top.$$

Thus, $x_1 = \sigma_1 V_1 \times (U^\top)_{11} + \sigma_2 V_2 \times (U^\top)_{21} + \cdots + \sigma_k V_k \times (U^\top)_{p1}$, which implies $(\sigma_1 (U^\top)_{11}, \cdots, \sigma_p (U^\top)_{p1})$ is a representation of x_1 with respect to $(V_1, \cdots, V_p)$.

SVD and representation

For $i = 1, \cdots, n$,

$$x_i = \sigma_1 (U^\top)_{1i} \times V_1 + \sigma_2 (U^\top)_{2i} \times V_2 + \cdots + \sigma_p (U^\top)_{pi} \times V_p$$

and $(U^\top)_{ji} = U_{ij}$, the following interpretations are derived from SVD.

- New (orthonormal) coordinate system: $(V_1, \cdots, V_p)$
- Scaling factor: $(\sigma_1, \cdots, \sigma_p)$
- Representation of x_i w.r.t the $(V_1, \cdots, V_p)$: $(\sigma_1 U_{i1}, \cdots, \sigma_p U_{ip})$.

Singular Value Decomposition: Example I

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

There is three observation in $\mathbb{R}^2$. $x_1^\top = (3, 1)$, the representation of the ith obs with respect to $(1, 0)^\top, (0, 1)^\top$.

$$\mathbf{U} = \begin{bmatrix} -0.5026 & 0.7746 \\ -0.5740 & -0.6325 \\ -0.6464 & 0.0000 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

The first row of $\mathbf{U}$, $(-0.5026, 0.7746)$ is the representation of x_1 with respect to $(-0.7071, -0.7071)$ and $(-0.7071, 0.7071)$.

Rank-k Truncated SVD: Definition & Structure

Full SVD (rank p)

$$\mathbf{X} = \sum_{i=1}^{p} \sigma_i \, U_i V_i^\top, \quad \sigma_1 \geq \cdots \geq \sigma_p > 0.$$

If $\sigma_j \simeq 0$ for all $j > k$, then $\mathbf{X} \simeq \sum_{i=1}^{k} \sigma_i \, U_i V_i^\top$. That is, x_i is represented based on the basis $\{V_1, \cdots, V_k\}$ of a k-dimensional subspace and $(U_{i1}, \cdots, U_{ik}) \in \mathbb{R}^k$ is the rank-reduced representation of x_i.

Rank-2 Truncated SVD: Visualization

Let $\mathbf{X}_k = \sum_{i=1}^{k} \sigma_i U_i V_i^\top$. and the denote ith row vector of $\mathbf{X}_k$ by $\tilde{x}_i^\top$.

- Axis: V_1 (horizontal) and V_2 (vertical)
- Interpretations of the Axis:
  $V_1 = (v_{11}, \cdots, v_{1p})^\top = \sum_{j=1}^{p} v_{1j} e_j$ and $V_2 = (v_{21}, \cdots, v_{2p})^\top = \sum_{j=1}^{p} v_{1j} e_j$,
  where e_js are the standard basis. V_j is explained by the covariates's names of the data and the associated coefficients $(v_{j1}, \cdots, v_{jp})$.
  (ex) Suppose that $V_1 = (0.7101, -0.7101, 0, \cdots, 0)$, X_1: GDP, X_2 : interest rate, then V_1 is the weighted sum of GDP and interest rate with the weight $(0.7101, -0.7101)$.
- Poisiton of $\tilde{x}_i^\top$: $(\sigma_1 U_{i1}, \sigma_2 U_{i2})$.

Principal Component Analysis

Concept of Dimensionality Reduction

- In many datasets, features are high-dimensional (e.g., image pixels, gene expressions).
- However, not all features contribute equally to the variation in data.
- **Dimensionality Reduction** aims to:
  - Eliminate redundant or noisy features
  - Find a low-dimensional representation of the data
  - Preserve the most informative aspects (variance or structure)
- Common motivations:
  - Visualization of high-dimensional data (2D/3D)
  - Faster computation and training
  - Mitigation of the “curse of dimensionality”

Examples of Dimensionality Reduction

- **Image Compression:** Represent high-resolution images (e.g., 1024x1024) using fewer principal components.
- **Text Data (NLP):** Reduce dimensionality of bag-of-words or TF-IDF vectors using techniques like Latent Semantic Analysis (LSA).
- **Gene Expression Data:** Microarray data with thousands of genes $\rightarrow$ identify a few latent variables (e.g., pathways).
- **Sensor Networks:** Combine redundant sensor readings into fewer representative signals.

Example: Gene Expression Data

- Genes are segments of DNA that code for proteins.
- Gene expression measures how actively a gene is being transcribed into RNA and translated into proteins.
- High expression = the gene is actively producing proteins.
  Low expression = little or no protein is being produced.
- Techniques: Microarray, RNA-seq provide expression levels as numeric values.

Gene Expression Dataset Structure

- Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the data matrix:
  - n: number of samples (e.g., tissues or patients)
  - p: number of genes (e.g., 20,000)
  - $(X)_{ij}$: expression level of gene j in sample i
- Each row $x_i^\top$: one sample
  Each column X_j: one gene

Centering the Data

- Remove gene-wise means (centering):

$$\tilde{\mathbf{X}} = \mathbf{X} - \Pi_{\mathcal{C}(1_n)}\mathbf{X}, \text{ where } \Pi_{\mathcal{C}(1_n)} = 1_n(1_n^\top 1_n)^{-1}1_n^\top$$

  Note that $\Pi_{\mathcal{C}(1_n)}$ is the projection matrix onto $1_n \in \mathbb{R}^n$.
- Result: each gene has mean 0 across samples
- This step is essential for PCA

Covariance Matrix

- Compute the gene-gene covariance matrix:

$$\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \in \mathbb{R}^{p\times p}$$

  Note that $(\mathbf{S})_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(\tilde{\mathbf{X}}^\top)_{ik}(\tilde{\mathbf{X}})_{kj} = \frac{1}{n-1}\sum_{k=1}^{n}(\tilde{\mathbf{X}})_{ki}(\tilde{\mathbf{X}})_{kj}$ Thus, $(\mathbf{S})_{ij}$ is the sample covrariance of X_i and X_j.
- Each entry S_{jk}: covariance between gene j and gene k
- Large values indicate strong co-expression

Principal Component Analysis (PCA)

By SVD $\tilde{X} = U^\top \Sigma V$ and $S = \frac{1}{n-1}\tilde{X}^\top \tilde{X} = V(\frac{1}{n-1}\Sigma^2)V^\top$

- Eigen-decomposition gives the same result

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

- $\mathbf{V}$: matrix of eigenvectors (principal directions)
- $\mathbf{\Lambda}$: diagonal matrix of eigenvalues (variances explained), $\mathbf{\Lambda} = \frac{1}{n-1}\Sigma^2$
- Choose top k components ($k = 2, 3$) in $\{(\lambda)_{ii} : i = 1, \cdots, p\}$ to reduce dimensionality.
- Principle Component Directions: $V_1, V_2, \cdots$

Projecting onto Principal Components

- Get sample representations in PC space: For the new coordinate with the largest eigenvalue (σ_1^2),

$$\mathbf{Z}_1 = \tilde{\mathbf{X}} V_1 \in \mathbb{R}^{n \times 1}$$

  Note that

$$\tilde{\mathbf{X}} V_1 = (\sum_{j=1}^{p} \sigma_j U_j V_j^\top) V_1 = U_j \in \mathbb{R}$$

- PC score of the jth PC direction: $\tilde{\mathbf{X}} V_j$ a sample's coordinate in a new coordinate V_j
- $\tilde{\mathbf{X}}[V_1 \ V_2] \in \mathbb{R}^{n \times 2}$ can be visualized to observe sample clustering or separation.

Example: Cancer vs Normal Samples

- Dataset: 100 samples (50 cancer, 50 normal), 20,000 genes
- PCA reveals clusters in 2D:
  - Cancer and normal samples may separate in the PC1-PC2 plane
  - PC1 may capture majority of variance due to disease status
- PC score in V_1 show genes contributing most to the variance
- See the Python Code! Click here!
- (HW) Image Compression?

Summary

- Gene expression data is high-dimensional and complex
- PCA provides an effective way to reduce dimensionality and visualize patterns
- Helps identify key genes, detect outliers, and classify samples

Matrix and Linear Map

Linear Map

Let A be $m \times n$ matrix and x be n matrix (n dimensional comlumn vector).

- Write an example of A and x and compute Ax. Where does the result lie on?
- Choose an other x' and compute Ax'.
- Choose two constant a and b and compute $A(ax)$ and $A(bx')$ and $A(ax) + A(bx')$.
- Compute $A(ax + bx')$.

Linear Map

- Write an example of A and x and compute Ax. Where does the result lie on?
  A moves $x \in \mathbb{R}^n$ on $Ax \in \mathbb{R}^m$.
- Choose an other x' and compute Ax'.
  A also moves $x' \in \mathbb{R}^n$ on $Ax' \in \mathbb{R}^m$.
- Choose two constant a and b and compute $A(ax)$ and $A(bx')$ and $A(ax) + A(bx')$.
- Compute $A(ax + bx')$.

Note that $A(ax) + A(bx') = A(ax + bx')$, which implies that A moves elements in $\mathbb{R}^n$ to $\mathbb{R}^n$ with satisfying an special property.

Definition: Linear map

Let V and W be vector spaces and let $\mathcal{L}$ be map from V to W.

- $\mathcal{L}(x+y) = \mathcal{L}(x) + \mathcal{L}(y)$ for all $x, y \in V$
- $\mathcal{L}(cx) = c\mathcal{L}(x)$ for a scalar c.

Matrix and linear map

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces, and consider a linear map $\mathcal{L}$ from $\mathcal{V}$ to $\mathcal{W}$. In particular, let $\mathcal{V} = \mathbb{R}^p$ and $\mathcal{W} = \mathbb{R}^n$, then $\mathcal{L}(\mathbf{0}) = \mathbf{0}$, and

$$\mathcal{L}(ax + bx') = a\mathcal{L}(x) + b\mathcal{L}(x')$$

for all $x, x' \in \mathbb{R}^p$ and all $a, b \in \mathbb{R}$.
Thus, $n \times p$ matrix can be regarded as a linear map. Moreover, we can consider one-to-one correspondence between linear map and matrix.

Matrix and linear map

- Matrix addition: let A and B be $n \times p$ matrix, and denote the corresponding linear map by $\mathcal{L}_A$ and $\mathcal{L}_B$. $A + B$ is also $n \times p$ matrix and $\mathcal{L}_{A+B}$ be the correspondent linear map to $A + B$. Then, $\mathcal{L}_{A+B} = \mathcal{L}_A + \mathcal{L}_B$.

$$(A + B)x = Ax + Bx$$

Matrix and linear map

- Matrix multiplication: let A and B be $n \times k$ and $k \times p$ matrix, and denote the corresponding linear map by $\mathcal{L}_A$ and $\mathcal{L}_B$. AB is $n \times p$ matrix and $\mathcal{L}_{AB}$ be the correspondent linear map to AB. Then, $\mathcal{L}_{AB} = \mathcal{L}_A \circ \mathcal{L}_B$ (Composition of functions)

$$x \mapsto Ax \mapsto B(Ax)$$

Matrix and linear map

Our conclusion is that

$$W \in \mathbb{R}^{n \times p} \text{ if and only if } W : \mathbb{R}^p \mapsto \mathbb{R}^n \text{ is linear.}$$

- When $n < p$ W is called a (linear) encoder.
- When $n > p$ W is called a (linear) decoder.

Moreover, it is also valid argument that the decompision of a matrix is that of the associated linear map. Next, we discuss a decomposition of squared matrix and will argue that it is the decomposition of the linear map.

Eigen-decomposition

Let $A \in \mathbb{R}^{p \times p}$ be a symmetric matrix. Then there exists an orthogonal matrix E and a diagonal matrix D (with real-valued elements) such that

$$A = EDE^\top$$

- Orthogonality of E: write

$$E = [e_1, \cdots, e_p]$$

  then $e_j^\top e_k = 0$ for $j \neq k$ and $\|e_j\| = 1$ for all j.
- Projection onto $\mathcal{C}(e_j)$ is given by $e_j(e_j^\top e_j)^{-1}e_j^\top = e_j e_j^\top$

Eigen-decomposition

Suppose that A be a symmetric matrix. Let λ_j be the jth diagonal element of D, then we can write

$$A = EDE^\top = \sum_{j=1}^{p} \lambda_j e_j e_j^\top$$

We can know that A is the sum of orthogonal projection operators. e_js are eigenvector and λ_j is the associated eigenvalue. $\mathcal{C}(e_j)$ is eigenspace spaned by e_j.

Eigen-decomposition

For simplicity let A be 2×2 matrix.

- Let $D_1 = \text{diag}(\lambda_1, 0)$ and $D_2 = \text{diag}(\lambda_2, 0)$, then

$$D_1 E^\top = \lambda_1 \begin{pmatrix} e_1^\top \\ 0 \end{pmatrix} \text{ and } D_2 E^\top = \lambda_2 \begin{pmatrix} 0 \\ e_2^\top \end{pmatrix}$$

- We can easily show that

$$\begin{pmatrix} e_1 & e_2 \end{pmatrix} \begin{pmatrix} e_1^\top \\ e_2^\top \end{pmatrix} = e_1 e_1^\top + e_2^\top e_2$$

Thus,

$$A = EDE^\top = E(D_1 E^\top + D_2 E^\top) = \lambda_1 e_1 e_1^\top + \lambda_2 e_2^\top e_2$$

Eigen-decomposition

This eigendecomposition can be viewed as the decomposition of a linear map:

$$\mathcal{L}_A = \sum_{j=1}^{p} \lambda_j \mathcal{L}_{E_j},$$

where $E_j = e_j e_j^\top$.
Note that

- projection onto $\mathcal{C}(e_j)$ is given by $e_j(e_j^\top e_j)^{-1} e_j^\top = e_j e_j^\top$

Eigen-decomposition

Therefore,

$$\mathcal{L}_A(x) = \sum_{j=1}^{p} \lambda_j \mathcal{L}_{E_j}(x),$$

where $\mathcal{L}_{E_j}(x)$ is projection onto the jth eigenspace.

Approximation of Linear map

Let $A^{(k)} = \sum_{j=1}^{k} \lambda_j e_j e_j^\top$ then $A^{(k)}$ approximates A?

Approximation of Linear map

$$
\begin{aligned}
EDE^\top \mathbf{x} &= [e_1, \cdots, e_p]\text{diag}(\lambda_1, \cdots, \lambda_p) \begin{pmatrix} e_1^\top \\ \vdots \\ e_p^\top \end{pmatrix} \mathbf{x} \\
&= [e_1, \cdots, e_p]\text{diag}(\lambda_1, \cdots, \lambda_p) \begin{pmatrix} e_1^\top \mathbf{x} \\ \vdots \\ e_p^\top \mathbf{x} \end{pmatrix} \\
&= [e_1, \cdots, e_p] \begin{pmatrix} \lambda_1 e_1^\top \mathbf{x} \\ \vdots \\ \lambda_p e_p^\top \mathbf{x} \end{pmatrix} \\
&= \sum_{j=1}^{p} e_j(\lambda_j e_j^\top \mathbf{x}) = (\sum_{j=1}^{p} \lambda_j e_j e_j^\top)\mathbf{x},
\end{aligned}
$$

Approximation of Linear map

Eigendecomposition shows the linear map of a symmetric matrix as the composition of three operations:

$$Ax = EDE^\top x$$

$$x \mapsto E^\top x \text{ (rotation)} \mapsto D(E^\top x) \text{ (scaling)}$$
$$\mapsto E(DE^\top x) \text{ (reverse rotation)}$$

Inverse matrix of positive definite matrix

Let $\mathbf{A}$ be symmetric and nonnegative definite matrix. Then the minimum eigenvalue is positive if and only if $\mathbf{A}$ is positive definite.

pf) Let λ_{min} be the minumum eigenvalue of $\mathbf{A}$. Assume that $\lambda_{min} > 0$. Let $\mathbf{x} = \sum_{j=1} a_j e_j \neq 0$, then

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{j=1}^{p} \lambda_j (e_j^\top \mathbf{x})^2 = \sum_{j=1}^{p} \lambda_j a_j^2 > 0.$$

Assume that $\mathbf{A}$ is pd matrix. WLOG, let λ_p be the minimum eigenvalue of $\mathbf{A}$. Then,

$$e_p^\top \mathbf{A} e_p = \sum_{j=1} \lambda_j (e_j^\top e_p)^2 = \lambda_p > 0.$$

Inverse matrix of positive definite matrix

The inverse matrix of such $\mathbf{A}$ is given by

$$\mathbf{A}^{-1} = \mathbf{E}D^{-1}\mathbf{E}^{\top}.$$

pf) $\mathbf{E}D^{-1}\mathbf{E}^{\top}\mathbf{A} = \mathbf{E}D^{-1}\underbrace{\mathbf{E}^{\top}\mathbf{E}}_{=I}D\mathbf{E}^{\top} = I$

and $\mathbf{A}\mathbf{E}D^{-1}\mathbf{E}^{\top} = \mathbf{E}D\underbrace{\mathbf{E}^{\top}\mathbf{E}}_{=I}D^{-1}\mathbf{E}^{\top} = I$. By definition of the inverse matrix, we obtain the result.

Summary

Suppose tha an asymmetic matrix A is decomposed by $A = E\Lambda E^\top$. Then

$$\mathcal{L}_A = \sum_{j=1}^{p} \mathcal{L}_{B_j},$$

where $B_j = e_j e_j^\top$ and e_j is the jth column vector of E.

See the Python Code! (Click here.)

Dimension Reduction and Optimal Reconstruction

Matrix Norms

Matrix norms measure the size or magnitude of a matrix. They play a crucial role in numerical analysis and matrix computations.

Commonly used matrix norms include:

- Operator Norm (Induced Norm)
- Frobenius Norm

Operator Norm

The operator norm (also called the induced norm) of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as:

$$\|\mathbf{A}\|_{\text{op}} = \sup_{x \neq \mathbf{0}} \frac{\|\mathbf{A}x\|_2}{\|x\|_2} = \sup_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}x\|_2$$

- Measures how much $\mathbf{A}$ stretches a vector.
- Equivalent to the largest singular value (i.e. σ_1 in SVD) of $\mathbf{A}$.
- Sub-multiplicative: $\|\mathbf{AB}\|_{\text{op}} \leq \|\mathbf{A}\|_{\text{op}} \|\mathbf{B}\|_{\text{op}}$

Frobenius Norm

The Frobenius norm of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as:

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |(A)_{ij}|^2 \right)^{1/2}$$

Alternatively,

$$\|A\|_F = \sqrt{\mathrm{Tr}(A^\top A)} = \left(\sum_{i=1}^{\min(m,n)} \sigma_i^2 \right)^{1/2}$$

- Equivalent to the Euclidean norm of the matrix as a vector.
- Easy to compute and differentiable.
- Unitary invariant: $\|\mathbf{UAV}\|_F = \|\mathbf{A}\|_F$ for orthogonal matrices $\mathbf{U}$, $\mathbf{V}$.

Linear Projection and Reconstruction

- Let each data point in high-dimensional space be denoted by $x_i \in \mathbb{R}^p$.
- Apply a linear transformation $\mathbf{W} \in \mathbb{R}^{p \times k}$ to obtain a low-dimensional representation:

$$z_i = \mathbf{W}^\top x_i \in \mathbb{R}^k$$

- We can reconstruct an approximation of the original data using a reconstruction matrix $\mathbf{W}_r \in \mathbb{R}^{p \times k}$:

$$\hat{x}_i = \mathbf{W}_r z_i = \mathbf{W}_r \mathbf{W}^\top x_i$$

- If $\mathbf{W}_r = \mathbf{W}$, this gives:

$$\hat{x}_i = \mathbf{W}\mathbf{W}^\top x_i$$

Example: Projection and Reconstruction

- Let $x_i = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \in \mathbb{R}^3$
- Define projection matrix $\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$
- Project to 2D:

$$z_i = \mathbf{W}^\top x_i = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

- Reconstruct:

$$\hat{x}_i = \mathbf{W} z_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$$

- Note: $\hat{x}_i \neq x_i \rightarrow$ Information loss

Optimal Linear Transformation for Fixed Subspace

- Given:
  - $x_i \in \mathbb{R}^p$ for $i = 1, \cdots, n$: original data. Assume that $\sum_{i=1}^n x_i = 0$.
  - $\mathbf{W} \in \mathbb{R}^{p \times k}$: fixed orthonormal basis (i.e., $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_k$)

- **Low-dimensional representation:** (Encoder)

$$z_i = f(x_i) \in \mathbb{R}^k$$

- **Reconstruction from low-dimensional representation:** (Decoder)

$$\hat{x}_i = \mathbf{W} z_i \in \mathbb{R}^p$$

- **Optimal Reconstruction**
  This choice minimizes the squared reconstruction error:

$$\min_f \sum_{i=1}^n \|x_i - \mathbf{W} f(x_i)\|^2$$

If decoder is linear, then optimal encoder is always linear. Also, the optimal encoder is completely determined by the given decoder.

Optimal Linear Transformation for Fixed Subspace

Surprisingly, $z_i = \mathbf{W}^\top x_i$ for any $\mathbf{W} \in \mathbb{R}^{p \times k}$

How to Choose the Best Subspace?

- So far, we assumed a fixed subspace defined by $\mathbf{W}$.
- Now we ask: **How do we choose the optimal subspace $\mathbf{W}$ itself?**
- Goal:
  - Find $\mathbf{W} \in \mathbb{R}^{p \times k}$ (with orthonormal columns) that minimizes the total reconstruction error:

$$\min_{\mathbf{W}:\mathbf{W}^\top \mathbf{W} = I_k} \frac{1}{n} \sum_{i=1}^{n} \|x_i - \mathbf{W}\mathbf{W}^\top x_i\|^2$$

- This is equivalent to:
  - Finding the subspace that captures the **maximum variance** of the data.
- This leads to: **Principal Component Analysis (PCA)**

Principal Component Analysis (PCA)

- Let the data be mean-centered: $\frac{1}{n}\sum_{i=1}^{n} x_i = \mathbf{0}$
- Define the sample covariance matrix:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \in \mathbb{R}^{p\times p}$$

- Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the data matrix with row vectors $x_i^\top$.
- Note that $\mathbf{S} = \mathbf{X}^\top\mathbf{X}/n$

Reconstruction Error as Matrix Form (1/2)

- Since $\hat{x}_j^\top = x_j^\top WW^\top$, the entire reconstruction can be written as:

$$\widehat{\mathbf{X}} = \mathbf{X}\mathbf{W}\mathbf{W}^\top$$

- The total squared reconstruction error becomes:

$$\sum_{i=1}^{n} \left\| x_i - \mathbf{W}\mathbf{W}^\top x_i \right\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} (\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top)_{ij}^2,$$

  where $(A)_{ij}$ denote the element of the ith row and jth column.

- This uses the Frobenius norm:

$$\|\mathbf{A}\|_F^2 = \sum_{i,j} a_{ij}^2 = \mathrm{Tr}(\mathbf{A}^\top \mathbf{A})$$

Reconstruction Error as Matrix Form (2/2)

- Expand the squared Frobenius norm:

$$\left\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{W}^{\top}\right\|_{F}^{2}=\operatorname{Tr}\left[(\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{W}^{\top})^{\top}(\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{W}^{\top})\right]$$

- Algebraic simplification:

$$=\operatorname{Tr}\left(\mathbf{X}^{\top}\mathbf{X}-\mathbf{X}^{\top}\mathbf{X}\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}\mathbf{W}^{\top}\mathbf{X}^{\top}\mathbf{X}+\mathbf{W}\mathbf{W}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{W}\mathbf{W}^{\top}\right)$$

- Use symmetry ($\mathbf{X}^{\top}\mathbf{X}$ is symmetric) and orthonormality ($\mathbf{W}^{\top}\mathbf{W}=\mathbf{I}$):

$$=\operatorname{Tr}\left(\mathbf{X}^{\top}\mathbf{X}-\mathbf{W}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{W}\right)$$

- Therefore:

$$\min_{\mathbf{W}^{\top}\mathbf{W}=\mathbf{I}_{k}}\left\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{W}^{\top}\right\|_{F}^{2}\quad\Longleftrightarrow\quad\max_{\mathbf{W}^{\top}\mathbf{W}=\mathbf{I}_{k}}\operatorname{Tr}\left(\mathbf{W}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{W}\right)$$

PCA via Eigen Decomposition

- Define the sample covariance matrix:

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{p\times p}$$

- Then the PCA objective becomes the problem finding $\mathbf{W}$,

$$\max_{\mathbf{W}^\top \mathbf{W}=\mathbf{I}_k} \operatorname{Tr}\left(\mathbf{W}^\top \mathbf{S}\mathbf{W}\right)$$

- Solution:
  - Let $\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}^\top$ be the eigen-decomposition of $\mathbf{S}$
  - $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$
  - $\mathbf{V} = [V_1, V_2, \ldots, V_p]$ are the eigenvectors
- Optimal choice of $\mathbf{W} \in \mathbb{R}^{p\times k}$:

$$\mathbf{W} = [V_1, \ldots, V_k] \quad \text{(top-}k\text{ eigenvectors of } \mathbf{S}\text{)}$$

PCA Solution via Eigen Decomposition (Rank-1 Case)

- PCA objective (rank-1 case):

$$\max_{\|w\|_2=1} w^\top \mathbf{S} w$$

  where $w \in \mathbb{R}^p$ and $\mathbf{S} \in \mathbb{R}^{p \times p}$ is symmetric and positive semi-definite.
- Let $\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}^\top$ be the eigen-decomposition of $\mathbf{S}$:

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p] \text{ is orthonormal}$$

- By substituting $w = \mathbf{V}a$, the problem becomes:

$$\max_{\|a\|_2=1} a^\top \Lambda a = \sum_{i=1}^{p} \lambda_i a_i^2$$

- The maximum is achieved when $a = [1, 0, \ldots, 0]^\top \Rightarrow w = V_1$
- Therefore, the optimal direction is the first eigenvector:

$$w^* = V_1$$

PCA Solution via Eigen Decomposition (1/2)

- Recall the PCA objective:

$$\max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}_k} \operatorname{Tr}(\mathbf{W}^\top \mathbf{S} \mathbf{W})$$

  where $\mathbf{S} \in \mathbb{R}^{p \times p}$ is symmetric and positive semi-definite.

- Let $\mathbf{S} = \mathbf{V} \Lambda \mathbf{V}^\top$ be the eigen-decomposition of $\mathbf{S}$:

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p), \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

$$\mathbf{V} = [V_1, \ldots, V_p] \text{ is orthonormal}$$

PCA Solution via Eigen Decomposition (1/2)

- Define $\mathbf{U} := \mathbf{V}^\top \mathbf{W} \in \mathbb{R}^{p \times k}$, then $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k$
- Then:

$$\mathrm{Tr}(\mathbf{W}^\top \mathbf{S} \mathbf{W}) = \mathrm{Tr}(\mathbf{U}^\top \Lambda \mathbf{U})$$

- $Tr(UU^\top) = Tr(U^\top U) = Tr(I_k) = k$

PCA Solution via Eigen Decomposition (2/2)

- Recall:

$$\mathrm{Tr}(\mathbf{U}^\top \Lambda \mathbf{U}) = \sum_{i=1}^{p} \lambda_i \sum_{j=1}^{k} u_{ij}^2 \quad \text{with} \quad \sum_{i=1}^{p} u_{ij}^2 = 1, \quad \sum_{i=1}^{p} \sum_{j=1}^{k} u_{ij}^2 = k$$

- From the constraints, $\sum_{j=1}^{k} u_{ij}^2 \leq 1$ for $i = 1, \cdots, p$ (See the Appendix).
- Thus, to maximize the weighted sum $\sum_{i=1}^{p} \lambda_i \cdot \left(\sum_{j=1}^{k} u_{ij}^2 \right)$, we must assign weights to the largest λ_i's.

PCA Solution via Eigen Decomposition (2/2)

- Maximum trace is achieved when:

$$\mathbf{W} = [V_1, \ldots, V_k] \quad \text{(top-}k\text{ eigenvectors of } \mathbf{S})$$

$$\mathrm{Tr}(\mathbf{W}^\top \mathbf{S}\mathbf{W}) = \sum_{i=1}^{k} \lambda_i$$

- This shows that PCA selects the directions of maximum variance.

PCA Solution via Eigen Decomposition (2/2)

Our conclusion

- Goal:
  - Find $\mathbf{W} \in \mathbb{R}^{p \times k}$ (with orthonormal columns) that minimizes the total reconstruction error:

$$\min_{\mathbf{W}:\mathbf{W}^\top \mathbf{W} = I_k} \frac{1}{n} \sum_{i=1}^{n} \|x_i - \mathbf{W}\mathbf{W}^\top x_i\|^2$$

- This is equivalent to:
  - Finding the subspace that captures the **maximum variance** of the data.

$$\mathbf{W} = [V_1, \ldots, V_k] \quad \text{(top-}k\text{ eigenvectors of } \mathbf{S}\text{)}$$

Appendix

Row–norm bound for orthonormal-column matrix

Claim

Let $\mathbf{U} \in \mathbb{R}^{p \times k}$ satisfy $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k$. For each row vector $u_i^\top (= (u_{i1}, \ldots, u_{ik}))$,

$$\|u_i\|_2^2 = \sum_{j=1}^{k} u_{ij}^2 \leq 1 \quad (i = 1, \ldots, p).$$

Auxiliary facts

- Column orthonormality implies $\sum_{i=1}^{p} u_{ij}^2 = 1$ for every j.
- Hence the total squared sum is $\sum_{i=1}^{p} \sum_{j=1}^{k} u_{ij}^2 = k$.

Proof of the row–norm bound I

1. **Orthogonal projector.** Define $\mathbf{P} := \mathbf{U}\mathbf{U}^\top \in \mathbb{R}^{p\times p}$. Because $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$,

$$\mathbf{P}^\top = \mathbf{P}, \qquad \mathbf{P}^2 = \mathbf{P}.$$

   Thus $\mathbf{P}$ is a symmetric projector whose eigenvalues are 1 (multiplicity k) and 0 (multiplicity $p-k$). Consequently $\|\mathbf{P}\|_2 = 1$.

2. **Diagonal entries of $\mathbf{P}$.** For the ith standard basis vector e_i,

$$P_{ii} = e_i^\top \mathbf{P} e_i = e_i^\top \mathbf{U}\mathbf{U}^\top e_i = \|\mathbf{U}^\top e_i\|^2 = \sum_{j=1}^{k} u_{ij}^2.$$

Proof of the row–norm bound II

3. **Eigenvalue of $\mathbf{P}$.** Since $\mathbf{P}^2 = \mathbf{P}$, $\mathbf{P}^2 v = \mathbf{P} v = \lambda v$. for all v. Let v be an eigenvector of $\mathbf{P}$ then $\mathbf{P}^2 v = \mathbf{P}\mathbf{P} v = \mathbf{P}(\lambda v) = \lambda(\mathbf{P} v) = \lambda^2 v$. Therefore

$$\lambda^2 \mathbf{v} = \lambda \mathbf{v} \quad \Longrightarrow \quad (\lambda^2 - \lambda)\mathbf{v} = \mathbf{0}.$$

   Because $\mathbf{v} \neq \mathbf{0}$, we must have

$$\lambda^2 - \lambda = 0 \quad \Longrightarrow \quad \lambda \in \{0, 1\}.$$

4. **Spectral bound.** For any unit vector $\mathbf{v}$, $\mathbf{v}^\top \mathbf{P} \mathbf{v} \leq \|\mathbf{P}\|_2 \|\mathbf{v}\|_2^2 = 1$. Taking $\mathbf{v} = \mathbf{e}_i$ gives $P_{ii} \leq 1$.
5. **Conclusion.** Combining steps 2 and 3,

$$\sum_{j=1}^{k} u_{ij}^2 = P_{ii} \leq 1 \qquad (i = 1, \ldots, p).$$