Principal Component Analysis

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

March 27, 2023

References

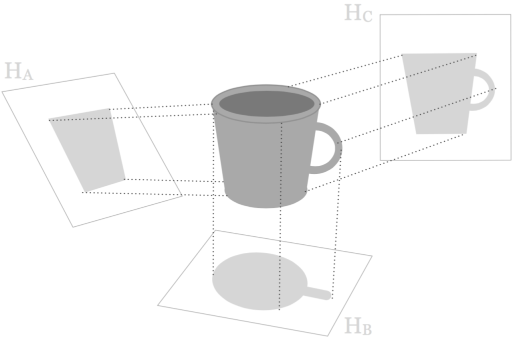
The contents of the slides are from: Gaston Sanchez and Ethan Marzban: *All Models Are Wrong: Concepts of Statistical Learning* - `https://allmodelsarewrong.github.io/pca.html`

Low-dimensional Representations

- Individuals form a cloud of points in a p-dim space. Variables form a cloud of arrows in an n-dim space.
- Suppose some data in which its cloud of points form a mug:

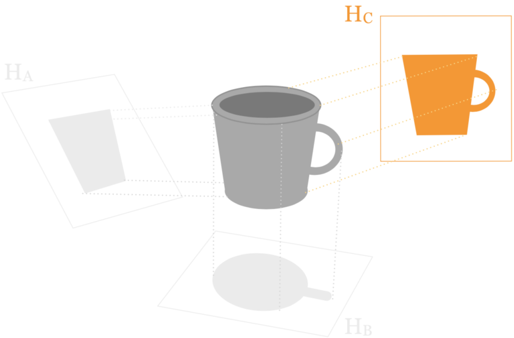
- Is there away to get a low-dimensional representation of this data?

Low-dimensional Representations



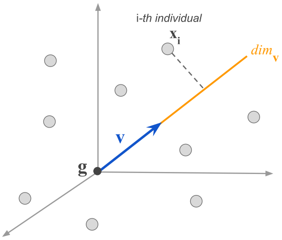
- We can look for projections of the data into sub-spaces of lower dimension.
- Assume we take a photo of the mug from different angles. What is the **best** angle to take a photo to get the images of the mug as similar as possible to the mug?

Low-dimensional Representations



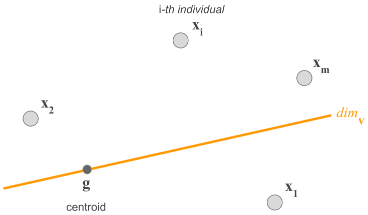
- Among 03 projections $\mathbb{H}_A, \mathbb{H}_B, \mathbb{H}_C$, the subspace $\mathbb{H}_C$ provides the best low-dimensional representation.
- The resulting image in low-dimensional space is not capturing the whole pattern: there is always some loss of information.
- Choosing the right projection, we try to minimize such loss.

Projections



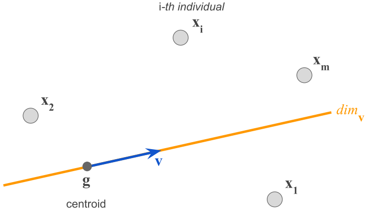
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.

Projections



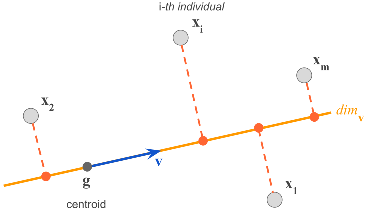
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.

Projections



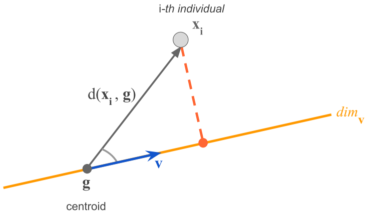
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.
- We manipulate $dim_{\mathbf{v}}$ via a vector $\mathbf{v}$ along this dimension.

Projections



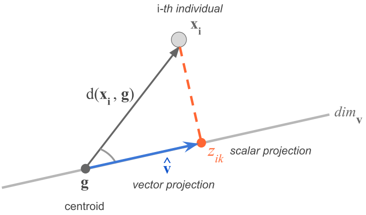
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.
- We manipulate $dim_{\mathbf{v}}$ via a vector $\mathbf{v}$ along this dimension.
- We want to project orthogonally the individuals onto this dimension.

Vector and Scalar Projections



- Take the centroid $\boldsymbol{g}$ as the origin of the clouds of points.
- The dimension that we look for has to pass through the origin.
- Obtain the orthogonal projection of the i-th individual onto $dim_{\mathbf{v}}$ is projecting $\mathbf{x}_i$ onto any vector $\mathbf{v}$ along this dimension.

Vector and Scalar Projections



- The **vector projection** of $\mathbf{x}_i$ onto $\mathbf{v}$ is:

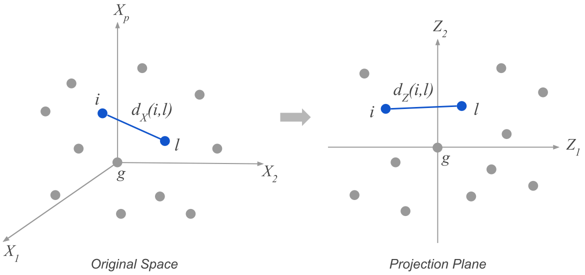
$$\hat{\mathbf{v}} = \frac{\mathbf{v}^\top \mathbf{x}_i}{\mathbf{v}^\top \mathbf{v}} \mathbf{v}$$

- The **scalar projection** of $\mathbf{x}_i$ onto $\mathbf{v}$ is:

$$z_{ik} = \frac{\mathbf{v}^\top \mathbf{x}_i}{\|\mathbf{v}\|}$$

- We would prefer the **scalar projection** to obtain the co-ordinate of $\mathbf{x}_i$ along this axis.

Projected Inertia



- Find the angle that give the best photo of the object $\iff$ Find the subspace that the distances between the points are the most similar to the original points.
- The overall dispersion of the original data is: $\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l)$. We try to find a subspace $\mathbb{H}$ such that:

$$\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l) \approx \sum_{i=1}^{n}\sum_{l=1}^{n} d_{\mathbb{H}}^2(i,l)$$

Projected Inertia

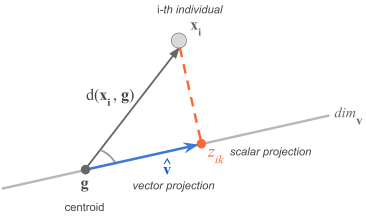
- The overall dispersion is related to the inertia as:

$$\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l) = 2n\sum_{i=1}^{n} d^2(i,g) = 2n^2\frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = 2n^2\texttt{Inertia}$$

- Finding the subspace $\mathbb{H}$ that yields similar distances to the original subspace corresponds to maximize the projected inertia:

$$\max_{\mathbb{H}}\left\{\frac{1}{n}\sum_{i=1}^{n} d_{\mathbb{H}}^2(i,g)\right\}$$

Projected Inertia



- We are consider $1D$ case, $\mathbb{H} \subseteq \mathbb{R}^1$, the projected inertia becomes:

$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathbb{H}}^2(i, g) = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{v})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2$$

- Our maximization problem becomes:

$$\max_{\mathbf{v}} \left\{ \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{v})^2 \right\} \quad \texttt{s.t.} \quad \mathbf{v}^\top \mathbf{v} = 1$$

- We constraint $\mathbf{v}$ to be a unit vector; otherwise, the maximization objective is unbounded.

Maximization Problem

- Assume mean-centered data, the centroid $\mathbf{g}$ of the cloud of points is the origin $\mathbf{g} = \mathbf{0}$.
- We are projecting onto a line spanned by a unit-vector $\mathbf{v}$, the projected inertia $I_{\mathbb{H}}$ is the variance of the projected data points:

$$
\begin{aligned}
I_{\mathbb{H}} &= \frac{1}{n}\sum_{i=1}^{n} d_{\mathbb{H}}^2(i,0) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{v})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\mathbf{z}^\top\mathbf{z} = \frac{1}{n}\mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}
\end{aligned}
$$

where

$$
\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \mathbf{X}\mathbf{v} = \begin{bmatrix} ---\mathbf{x}_1^\top --- \\ ---\mathbf{x}_2^\top --- \\ -------- \\ ---\mathbf{x}_n^\top --- \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{pmatrix}
$$

Maximization Problem

- The maximization problem becomes:

$$\max_{\mathbf{v}} \left\{ \frac{1}{n} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \right\} \quad \texttt{s.t.} \quad \mathbf{v}^\top \mathbf{v} = 1$$

- To solve this maximization, problem, we use Lagrange multipliers.

$$\mathcal{L} = \frac{1}{n} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} - \lambda(\mathbf{v}^\top \mathbf{v} - 1)$$

- Set the derivative of the Lagrangian $\mathcal{L}$ wrt $\mathbf{v}$ to $\mathbf{0}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = \frac{2}{n} \mathbf{X}^\top \mathbf{X} \mathbf{v} - 2\lambda \mathbf{v} = \mathbf{0} \Rightarrow \underbrace{\frac{1}{n} \mathbf{X}^\top \mathbf{X}}_{\mathbf{S} \in \mathbb{R}^{p \times p}} \mathbf{v} = \lambda \mathbf{v} \Rightarrow \mathbf{S} \mathbf{v} = \lambda \mathbf{v}$$

- This means that $\mathbf{v}$ is an eigenvector (with eigenvalue λ) of $\mathbf{S}$.
- λ is the value of the projected inertia $I_{\mathbb{H}}$ that we want to maximize.

Eigenvectors of $\mathbf{S}$

- Assume $\mathbf{X}$ is full rank ($rank(\mathbf{X}) = p$). We have p eigenvectors:

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_k & \dots & \mathbf{v}_p \end{bmatrix}$$

- We also have the matrix of eigenvalues $\mathbf{\Lambda} = \mathtt{diag}\{\lambda_i\}_{i=1}^{n}$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

- We then have the matrix of projected points $\mathbf{Z}$ (also known as **the matrix of principal components (PC's)**):

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_k & \dots & \mathbf{z}_p \end{bmatrix}$$

  where the k-th principal component $\mathbf{z}_k$ is:

$$\mathbf{z}_k = \mathbf{X}\mathbf{v}_k = v_{1k}\mathbf{x}_1 + v_{2k}\mathbf{x}_2 + \dots + v_{pk}\mathbf{x}_p$$

  with $\mathbf{x}_k$ denotes columns of $\mathbf{X}$.

Eigenvalues of $\mathbf{S}$

- Because the data is mean-centered, we have $\texttt{mean}(\mathbf{x}_i) = 0$. Then, $\texttt{mean}(\mathbf{z}_k) = 0$.
- How about the variance of $\mathbf{z}_k$?

$$
\begin{aligned}
Var(\mathbf{z}_k) &= \frac{1}{n}\mathbf{z}^\top\mathbf{z} = \frac{1}{n}(\mathbf{X}\mathbf{v}_k)^\top(\mathbf{X}\mathbf{v}_k) = \frac{1}{n}\mathbf{v}_k^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}_k \\
&= \mathbf{v}_k^\top\mathbf{S}\mathbf{v}_k = \mathbf{v}_k^\top(\lambda_k\mathbf{v}_k) = \lambda_k(\mathbf{v}_k^\top\mathbf{v}_k) = \lambda_k
\end{aligned}
$$

- The k-th eigenvalue of $\mathbf{S}$ is the variance of the k-th principal component.
- If $\mathbf{X}$ is mean centered, $\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is the covariance matrix of data.
- If $\mathbf{X}$ is standardized (mean-centered and scaled by the variance), then $\mathbf{S}$ is the correlation matrix.

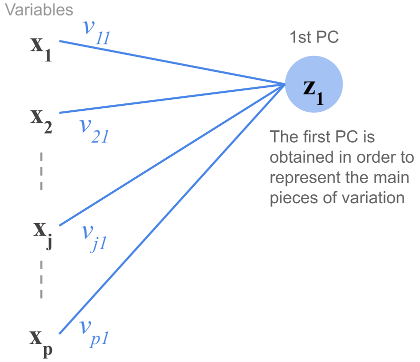
Eigenvalues of $\mathbf{S}$

$$\texttt{Inertia} = \frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = \sum_k \lambda_k = \texttt{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)$$

- $\sum_{k=1}^{p} \lambda_k$ relates to the total amount of variability in the data.
- The principal components capture different parts of the variability in the data.

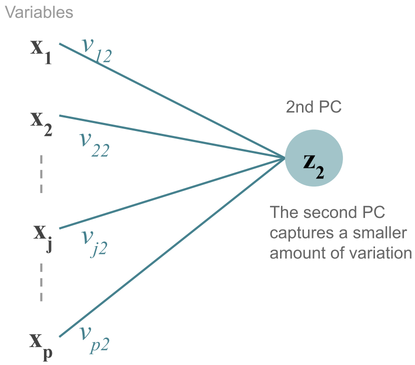
Principal Component Analysis (PCA)

- Given a set of p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, we want to obtain new k variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, called the **Principal Components (PCs)**.
- A principal component is a **linear combination** of the p variables: $\mathbf{z} = \mathbf{X}\mathbf{v}$.
- The first PC is a linear combination:



Principal Component Analysis (PCA)

- Given a set of p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, we want to obtain new k variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, called the **Principal Components (PCs)**.
- A principal component is a **linear combination** of the p variables: $\mathbf{z} = \mathbf{X}\mathbf{v}$.
- The second PC is another linear combination:



Principal Component Analysis (PCA)

- Given a set of p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, we want to obtain new k variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, called the **Principal Components (PCs)**.
- A principal component is a **linear combination** of the p variables: $\mathbf{z} = \mathbf{X}\mathbf{v}$.
- We compute PCs as linear combinations of original variables:

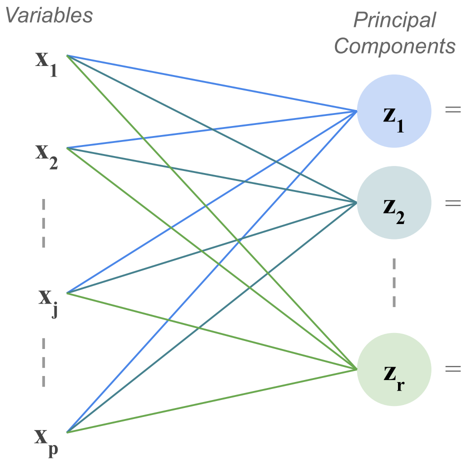
$$
\begin{aligned}
\mathbf{z}_1 &= v_{11}\mathbf{x}_1 + v_{21}\mathbf{x}_2 + \ldots + v_{p1}\mathbf{x}_p \\
\mathbf{z}_2 &= v_{12}\mathbf{x}_1 + v_{22}\mathbf{x}_2 + \ldots + v_{p2}\mathbf{x}_p \\
\vdots &= \vdots \\
\mathbf{z}_k &= v_{1k}\mathbf{x}_1 + v_{2k}\mathbf{x}_2 + \ldots + v_{pk}\mathbf{x}_p
\end{aligned}
$$

Or:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

where $\mathbf{Z}$ is an $n \times k$ matrix of principal components, and $\mathbf{V}$ is a $p \times k$ matrix of weights (directional vectors of the principal axes).

Principal Component Analysis (PCA)



$$= v_{11}\,\mathbf{x}_1 + v_{21}\,\mathbf{x}_2 + \ldots + v_{p1}\,\mathbf{x}_p$$

$$= v_{12}\,\mathbf{x}_1 + v_{22}\,\mathbf{x}_2 + \ldots + v_{p2}\,\mathbf{x}_p$$

$$= v_{1r}\,\mathbf{x}_1 + v_{2r}\,\mathbf{x}_2 + \ldots + v_{pr}\,\mathbf{x}_p$$

Finding Principal Components

- The components $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$ are required to capture most of the variation in data $\mathbf{X}$.
- We look for a vector $\mathbf{v}_h$ such that a component $\mathbf{z}_h = \mathbf{X}\mathbf{v}_h$ has maximum variance:

$$\max_{\mathbf{v}_h} var(\mathbf{z}_h) \Rightarrow \max_{\mathbf{v}_h} var(\mathbf{X}\mathbf{v}_h) \Rightarrow \max_{\mathbf{v}_h} \frac{1}{n}\mathbf{v}_h^\top \mathbf{X}^\top \mathbf{X}\mathbf{v}_h$$

- If $\mathbf{v}_h$ can be arbitrarily big, the problem is unbounded. We need to restrict $\mathbf{v}_h$ to be of unit norm:

$$\|\mathbf{v}_h\| = 1 \Rightarrow \mathbf{v}_h^\top \mathbf{v}_h = 1$$

- If we denote the covariance matrix $\mathbf{S} = (1/n)\mathbf{X}^\top \mathbf{X}$, then

$$\max_{\mathbf{v}_h} \mathbf{v}_h^\top \mathbf{S}\mathbf{v}_h \quad \texttt{s.t.} \quad \mathbf{v}_h^\top \mathbf{v}_h = 1$$

- To avoid redundancy, we require $\mathbf{z}_h^\top \mathbf{z}_l = 0$ mutually orthogonal if $h \neq l$.

Finding Principal Components

All PCs can be found by **diagonalizing** $\mathbf{S} = (1/n)\mathbf{X}^\top\mathbf{X}$.

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

- Λ is a diagonal matrix. The diagonal elements of Λ are the eigenvalues of $\mathbf{S}$.
- The columns of $\mathbf{V}$ are orthonormal: $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$
- The columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{S}$.
- $\mathbf{V}^\top = \mathbf{V}^{-1}$

Because $\mathbf{S}$ is a $p \times p$ symmetric matrix, we have:

- $\mathbf{S}$ has p real eigenvalues.
- The eigenvectors corresponding to different eigenvalues are orthogonal. $\mathbf{S}$ is orthogonally diagonalizable ($\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$).
- The set of eigenvalues of $\mathbf{S}$ is called the **spectrum of $\mathbf{S}$**.
- The PCA is obtained via an Eigenvalue Decomposition of $\mathbf{S}$.

Examples

- Principal Component Analysis - Intuitions: https://fmin.xyz/docs/applications/pca/
- Principal Component Analysis - Explained Visually: https://setosa.io/ev/principal-component-analysis/
- Principal Component Analysis (PCA): Iris data: https://www.math.umd.edu/~petersd/666/html/iris_pca.html
- Face Recognition using Principal Component Analysis: https://machinelearningmastery.com/face-recognition-using-principal-component-analysis/