Principal Component Analysis

Mingmin Chi

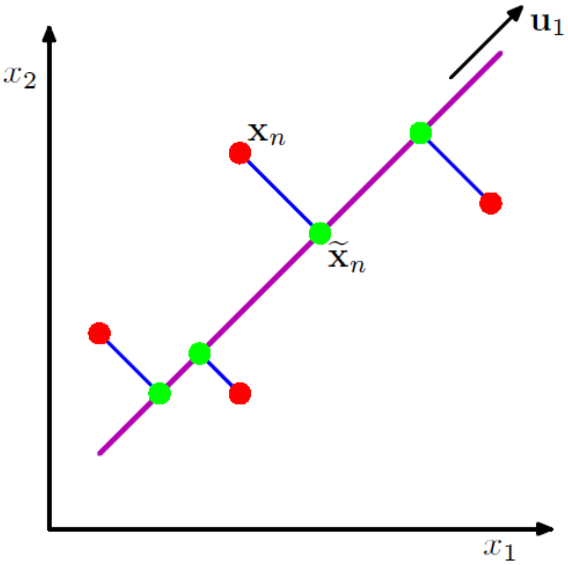
CSE Fudan University, Shanghai, China

- PCA is a technique widely used for dimensionality reduction, lossy data compression, feature extraction, and data visualization (Jolliffe, 2002)
- also known as the Karhunen-Loève (KL) transform

Definitions of PCA

- Maximum variance: orthogonal projection of the data onto a lower dimensional linear space (principal subspace) $\rightarrow$ the variance of the projected data is maximized (Hotelling, 1933)
- linear projection that minimizes the average projection cost (mean squared distance between the data points and their projections) (Pearson, 1901)

Maximum variance and minimum error



One-dimensional Projection

- Projecting the data onto a one-dimensional space ($M = 1, M < D$)
- Defining the direction of this space using a D-dimensional vector $\mathbf{v}_1$ and $\mathbf{v}_1^\top \mathbf{v}_1 = 1$
- the mean of the projected data is $\mathbf{v}_1^\top \boldsymbol{\mu}$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

The variance of the projected data is given by

$$\frac{1}{N}\sum_{i=1}^{n}\{\mathbf{v}_1^\top \mathbf{x}_i - \mathbf{v}_1^\top \boldsymbol{\mu}\}^2 = \mathbf{v}_1^\top \mathcal{C} \mathbf{v}_1$$

where $\mathcal{C}$ so-called *scatter matrix* or *covariance matrix* is defined by $\mathcal{C} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$

One-dimensional Projection

With the constraint, we have

$$\begin{aligned} &\mathbf{v}_1^\top \mathcal{C}\mathbf{v}_1 + \lambda_1(1 - \mathbf{v}_1^\top \mathbf{v}_1) \\ &\mathcal{C}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \\ \equiv &\mathbf{v}^\top \mathcal{C}\mathbf{v} = \lambda \mathbf{v}^\top \mathbf{v} = \lambda \end{aligned}$$

- to maximize $\mathbf{v}^\top \mathcal{C}\mathbf{v} \rightarrow$ to select the eigenvector corresponding to the largest eigenvalue of the scatter matrix
- this eigenvector is known as the first principal component

PCA - Multi-dimensional Projection

Additional principal components can be selected to maximize the projected variance amongst all possible directions orthogonal to those already considered, i.e., $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M$

An alternative formulation of PCA based on projection error minimization

Introducing a complete orthonormal set of D-dimensional basis vectors $\mathbf{u}_i$, where $i = 1, \cdots, D$ that satisfy

$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

Since this basis is complete, each data point can be represented by a linear combination of the basis vectors

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni} \mathbf{u}_i$$

Coordinate Rotation

- A rotation of coordinate system from the original $\mathbf{x}$ to a new system defined by the $\mathbf{u}_i$
- the original D components are replaced by an equivalent set $\alpha_{n1}, \cdots, \alpha_D$, where

$$\alpha_{ni} = \mathbf{x}_n^\top \mathbf{u}_i$$

Therefore,

$$\mathbf{x}_n = \sum_{i=1}^{D} (\mathbf{x}_n^\top \mathbf{u}_i)\mathbf{u}_i$$

Approximate Representation

Our goal is to approximate the data point using a representation involving a restricted number $M < D$ of variables corresponding to a projection onto a lower-dimensional subspace

- the M-dimensional linear subspace can be represented by the first M of the basis vectors
- Approximating each data point by

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} b_i\mathbf{u}_i$$

with the constraint that the coefficients b_i are constant

Minimizing Approximate Error

The goal is to minimize the approximate error by the reduction of dimensionality, i.e., the selection of $\mathbf{u}_i, z_{ni}, b_i$ by minimizing

$$J = \frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}_n - \tilde{\mathbf{x}}_n||^2$$

- Obtaining $z_{ni} = \mathbf{x}_n^\top \mathbf{u}_i, i = 1, \cdots, M$: setting the derivative w.r.t. z_{ni} to zero
- Obtaining $b_i = \boldsymbol{\mu}^\top \mathbf{u}_i, i = M+1, \cdots, D$

Therefore, we have the distortion measure in the form

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D} (\mathbf{x}_n^\top \mathbf{u}_i - \boldsymbol{\mu}^\top \mathbf{u}_i)^2$$

$$= \sum_{i=M+1}^{D} \mathbf{u}_i^\top \mathcal{C} \mathbf{u}_i$$

Minimum Measure

Since $\mathbf{u}_i \top \mathbf{u}_i = 1$, we have the similar result as minimum variance case by solving

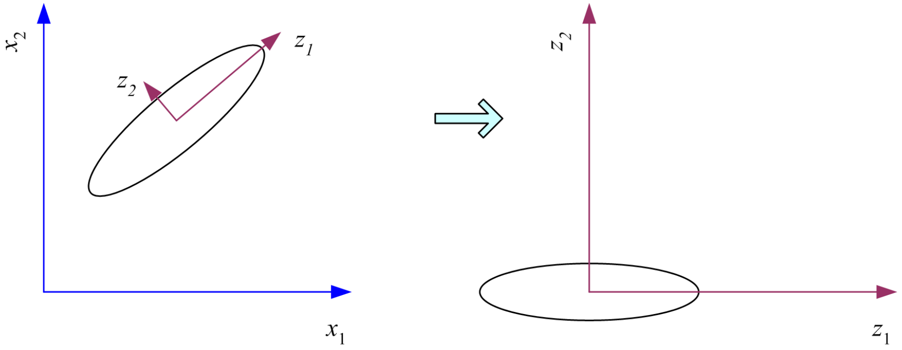
$$\mathcal{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

and to minimize $\mathbf{u}_i^\top \mathcal{C} \mathbf{u}_i$ by selecting

$$J = \sum_{i=M+1}^{D} \lambda_i$$

Visualization

centers the data at the origin and rotates the axes



Algorithm for PCA

(1) Computing the d-dimensional mean $\boldsymbol{\mu}$ and $D \times D$ covariance matrix $\Sigma = \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top$ for the full dataset

(2) Calculating the eigenvalues and the corresponding eigenvectors in terms of the eigenvalue function

$$\Sigma\mathbf{u} = \lambda\mathbf{u}$$

(3) Sorting the eigenvalues according to decreasing order

(4) Choosing the M largest eigenvalues and the corresponding eigenvectors according to

$$\frac{\sum_{i=1}^{D}\lambda_i}{\sum_{j=1}^{D}\lambda_j} \geq \tau, \ \tau < 1$$

Lower dimensional representation by PCA

(5) Constructing $D \times M$ matrix $\mathbf{A}$ whose columns consist of the M eigenvectors

The representation of data by principal components consists of projecting the data onto the M-dimensional subspace according to

$$\tilde{\mathbf{x}} = \mathbf{A}^\top(\mathbf{x} - \boldsymbol{\mu})$$

Conclusion

- The goal of PCA is to find a set of orthogonal components that minimize the error in the reconstructed data. An equivalent formulation of PCA is to find an orthogonal set of vectors that maximize the variance of the projected data.
- In other words, PCA seeks a transformation of the data into another frame of reference with as little error as possible, using fewer factors than the original data.
- For example, people often use PCA to reduce the dimensionality of data, that is, transforming d sensor readings into a set of p important factors in those readings.