Principal Component Analysis

Dit-Yan Yeung

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

COMP 4211: Machine Learning (Fall 2022)

1. Dimensionality Reduction
2. Principal Component Analysis
3. Some Practical Issues
4. Further Study

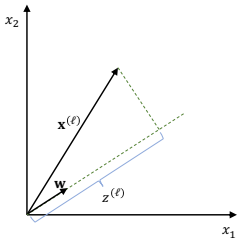
Dimensionality Reduction

- Dimensionality reduction refers to the process of reducing the number of variables or dimensions to be considered by the subsequent machine learning model (such as classification).
- Having data of high dimensionality usually implies:
  - High computational cost for learning and inference
  - Large number of model parameters with a higher chance of overfitting.
- Reducing the data dimensionality also makes it easier for data visualization and interpretation.

Feature Selection vs. Feature Transformation

- Feature selection:
  - From the d original features, choose a subset of $k\,(< d)$ features and discard the remaining $d - k$.
- Feature transformation or feature projection:
  - Find a mapping to project from the original d dimensions to $k\,(< d)$ dimensions.
  - While unsupervised methods do not use output information, supervised methods use it to find the mapping for feature transformation.
  - There exist both linear and nonlinear feature transformation methods. Linear ones are generally more efficient although nonlinear ones may be more effective.
- In this topic, we consider a linear, unsupervised feature transformation method.

Notation



- Let $\mathbf{x}$ denote a d-dimensional random vector, i.e., a vector of d random variables, corresponding to d input features.
- We assume that each example in a finite data set $\mathcal{S} = \{\mathbf{x}^{(\ell)}\}_{\ell=1}^{N}$ is a specific realization or observation of $\mathbf{x}$ sampled according to some (unknown) data distribution.
- The projection of an example $\mathbf{x}^{(\ell)}$ onto a unit vector $\mathbf{w} \in \mathbb{R}^d$ refers to the component of $\mathbf{x}^{(\ell)}$ in the direction of $\mathbf{w}$:

$$z^{(\ell)} = \mathbf{w}^\top \mathbf{x}^{(\ell)}.$$

Principal Component Analysis

- Principal component analysis (PCA) is an unsupervised dimensionality reduction method which finds a linear projection from the original d-dimensional input space to a new k-dimensional space ($k < d$) with minimum loss of data variance (or, equivalently, maximum preservation of data variance).
- Implicitly, PCA assumes that data variance is highly correlated with useful information that should be preserved.
- We use $\text{Var}(\cdot)$ to denote the population variance of a random variable, e.g., $\text{Var}(z)$ which is approximated by the sample variance computed based on $\mathcal{S}$.
- Since $z = \mathbf{w}^\top \mathbf{x}$, the data variance after projection onto $\mathbf{w}$ is $\text{Var}(z) = \text{Var}(\mathbf{w}^\top \mathbf{x})$ which varies with $\mathbf{w}$. To preserve as much data variance as possible is equivalent to finding $\mathbf{w}$ that maximizes $\text{Var}(z)$.

Data Variance

- Let $\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}]$ and $\mathbf{C} = \mathrm{E}\big[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\big]$ denote the mean vector and covariance matrix of $\mathbf{x}$, respectively.
- We can express $\mathrm{Var}(z)$ in terms of $\mathbf{C}$ and $\mathbf{w}$:

$$\begin{aligned}
\mathrm{Var}(z) &= \mathrm{Var}(\mathbf{w}^\top \mathbf{x}) \\
&= \mathrm{E}\big[\big(\mathbf{w}^\top \mathbf{x} - \mathrm{E}[\mathbf{w}^\top \mathbf{x}]\big)^2\big] \\
&= \mathrm{E}\big[(\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \boldsymbol{\mu})(\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \boldsymbol{\mu})\big] \\
&= \mathrm{E}\big[(\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \boldsymbol{\mu})(\mathbf{x}^\top \mathbf{w} - \boldsymbol{\mu}^\top \mathbf{w})\big] \\
&= \mathrm{E}\big[\mathbf{w}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{w}\big] \\
&= \mathbf{w}^\top \mathrm{E}\big[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\big] \mathbf{w} \\
&= \mathbf{w}^\top \mathbf{C}\, \mathbf{w}.
\end{aligned}$$

First Principal Component

- Optimization problem for finding the best projection direction $\mathbf{w}$:

$$\max_{\mathbf{w}} \text{Var}(\mathbf{w}^\top \mathbf{x}), \quad \text{subject to } \|\mathbf{w}\| = 1 \quad (\text{or, equivalently, } \mathbf{w}^\top \mathbf{w} = 1).$$

- This is a constrained optimization problem with an equality constraint which can be solved by introducing a Lagrange multiplier to define the Lagrangian for maximization:

$$L(\mathbf{w}, \alpha) \equiv \mathbf{w}^\top \mathbf{C}\, \mathbf{w} - \alpha(\mathbf{w}^\top \mathbf{w} - 1).$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$ and setting the derivative to $\mathbf{0}$, we get the eigenvalue equation for $\mathbf{w}$:

$$\mathbf{C}\, \mathbf{w} = \alpha\, \mathbf{w}. \tag{1}$$

First Principal Component (2)

- In general, the eigenvalue equation has d solutions which correspond to d eigenvectors $\mathbf{w}$ and d eigenvalues α.
- Premultiplying both sides of the eigenvalue equation by $\mathbf{w}^\top$, we get

$$\mathbf{w}^\top \mathbf{C}\,\mathbf{w} = \alpha\,\mathbf{w}^\top \mathbf{w} = \alpha.$$

- Since we want to maximize $\mathbf{w}^\top \mathbf{C}\,\mathbf{w}$ $(= \mathrm{Var}(z))$, we choose $\mathbf{w}$ to be the eigenvector with the largest eigenvalue and use it for projecting $\mathbf{x}$.
- Typically $k > 1$, i.e., we project $\mathbf{x}$ onto more than one direction. Thus, for clarity, we denote the $\mathbf{w}$ found above by $\mathbf{w}_1$ and the z, called the first principal component, by z_1.

Second Principal Component

- With $\mathbf{w}_1$ fixed, we now want to find the second unit vector $\mathbf{w}$ such that projecting $\mathbf{x}$ onto $\mathbf{w}$ will explain the maximum proportion of the remaining variance.
- To ensure that $\mathbf{w}$ only explains the remaining variance not already explained by $\mathbf{w}_1$, we need to enforce that $\mathbf{w}$ is orthogonal to $\mathbf{w}_1$.
- Optimization problem for finding the best second projection direction $\mathbf{w}$:

$$\max_{\mathbf{w}} \text{Var}(\mathbf{w}^\top \mathbf{x}), \quad \text{subject to } \mathbf{w}^\top \mathbf{w} = 1 \text{ and } \mathbf{w}^\top \mathbf{w}_1 = 0.$$

- Lagrangian with two Lagrange multipliers for maximization:

$$L(\mathbf{w}, \alpha, \beta) \equiv \mathbf{w}^\top \mathbf{C}\, \mathbf{w} - \alpha(\mathbf{w}^\top \mathbf{w} - 1) - \beta(\mathbf{w}^\top \mathbf{w}_1 - 0).$$

Second Principal Component (2)

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$ and setting the derivative to $\mathbf{0}$, we get the following equation:

$$2\,\mathbf{C}\,\mathbf{w} - 2\,\alpha\,\mathbf{w} - \beta\,\mathbf{w}_1 = \mathbf{0}. \tag{2}$$

- Premultiplying this equation by $\mathbf{w}_1^\top$ gives

$$2\,\mathbf{w}_1^\top\mathbf{C}\,\mathbf{w} - 2\,\alpha\,\mathbf{w}_1^\top\mathbf{w} - \beta\,\mathbf{w}_1^\top\mathbf{w}_1 = 0.$$

- Note that $\mathbf{w}_1^\top\mathbf{w} = 0$ and $\mathbf{w}_1^\top\mathbf{w}_1 = 1$. Thus $2\,\mathbf{w}_1^\top\mathbf{C}\,\mathbf{w} = \beta$.
- Since $\mathbf{C}\,\mathbf{w}_1 = \alpha_1\,\mathbf{w}_1$,

$$\mathbf{w}_1^\top\mathbf{C}\,\mathbf{w} = \left(\mathbf{w}_1^\top\mathbf{C}\,\mathbf{w}\right)^\top = \mathbf{w}^\top\mathbf{C}\,\mathbf{w}_1 = \alpha_1\,\mathbf{w}^\top\mathbf{w}_1 = 0.$$

Hence $\beta = 0$, implying that the second Lagrange multiplier (for enforcing that $\mathbf{w}$ is orthogonal to $\mathbf{w}_1$) is not necessary (because it is already implicitly enforced by the eigenvalue equation).

Second Principal Component (3)

- Consequently, equation (2) can be rewritten as

$$\mathbf{C}\,\mathbf{w} = \alpha\,\mathbf{w},$$

  which is exactly the same as equation (1) used for finding the first principal component.
- Using the same argument as before, we want to find an eigenvector that has as large an eigenvalue as possible.
- However, since the eigenvector $\mathbf{w}_1$ with the largest eigenvalue has already been chosen, we choose the eigenvector with the second largest eigenvalue, denoted by $\mathbf{w}_2$, as the projection direction for computing the second principal component.

Beyond the First Two Principal Components

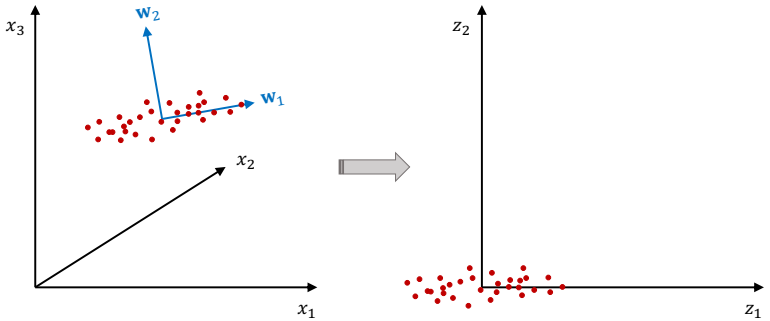
- To project $\mathbf{x}$ onto a k-dimensional space, we simply solve the eigenvalue equation (just once) and choose the k eigenvectors with the largest eigenvalues.
- From the perspective of linear algebra, the k eigenvectors are linearly independent basis vectors forming a basis in a vector space. The span of these basis vectors corresponds to the k-dimensional space onto which $\mathbf{x}$ is projected.
- In case two or more eigenvectors have the same eigenvalues, we can break the tie arbitrarily.

Linear Projection

- Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ be a $d \times k$ projection matrix formed by the k leading eigenvectors or basis vectors obtained when performing PCA on the data set $\mathcal{S}$.
- Let $\mathbf{m}$ be the sample mean of the N d-dimensional vectors in $\mathcal{S}$.
- Using $\mathbf{W}$ and $\mathbf{m}$, each $\mathbf{x}^{(\ell)}$ in $\mathcal{S}$ can be transformed into a k-dimensional vector in a new space:

$$\mathbf{z}^{(\ell)} = \mathbf{W}^\top (\mathbf{x}^{(\ell)} - \mathbf{m}).$$

Linear Projection (2)



Data Preprocessing

- If the variances of different original dimensions vary considerably, they may affect the resulting principal components more than the correlations between dimensions.
- Possible ways to alleviate this problem:
  - Preprocess the data so that each dimension has zero mean and unit variance before applying PCA.
  - Perform eigenvalue decomposition on the correlation matrix instead of the covariance matrix.

Computational Complexity

- The key computational step of PCA is to perform eigenvalue decomposition on a $d \times d$ sample covariance matrix.
- The worst-case computational complexity of eigenvalue decomposition is $\mathcal{O}(d^3)$.
- Faster methods exist for the special type of matrices in PCA, e.g., using the Coppersmith–Winograd algorithm with complexity $\mathcal{O}(d^{2.376})$.

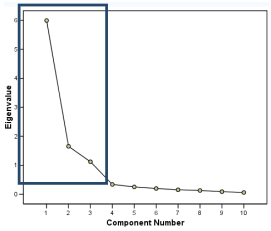
How to Choose k?

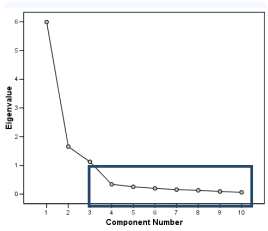
- Let us sort the eigenvalues in non-increasing order and denote them by $\lambda_1, \lambda_2, \ldots, \lambda_d$.
- Inspired by the analysis of variance (ANOVA) models in statistics, we can define the proportion of variance explained by the k leading principal components as follows:

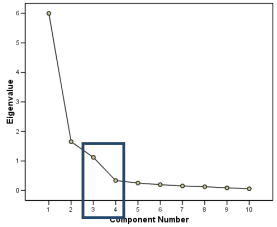
$$\text{PoV}(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}.$$

- We can set a threshold percentage θ (say, 90%) of the data variance to preserve and choose the smallest k such as $\text{PoV}(k) \geq \theta$.
- If the dimensions in the original input space are highly correlated, only a small number of eigenvectors will have large eigenvalues. Consequently, we will have $k \ll d$, implying a large reduction in dimensionality.

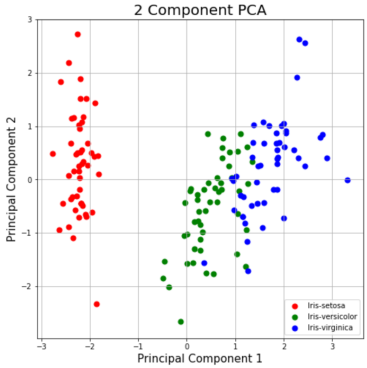
Scree Plot


Eigenvalue
Component Number


Eigenvalue
Component Number


Eigenvalue
Component Number

Data Visualization



- If PoV(2) is reasonably large, we may plot the data points using the two leading principal components to look for structure or grouping in the data set.

To Learn More...

- t-distributed stochastic neighbor embedding
- Nonlinear dimensionality reduction