Principal Component Analysis

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Dimensionality reduction

Goal of dimension reduction

Find a projection of the high dimensional data $\boldsymbol{x} \in \mathbb{R}^D$ to a low dimensional subspace $\boldsymbol{z} \in \mathbb{R}^L$ ($L < D$), such that the low dimensional representation is a "good approximation" to the original data.

$\boldsymbol{z}$ is called the **latent vector** because its values are latent or "hidden" or not observed in the data. The collection of these latent variables are called the **latent factors**.
Dimensionality reduction methods are used for :

- visualization
- data representation (noise reduction, pre-processing, hidden structure in data)

Model

For PCA, the projection is linear and orthogonal. We project or encode $\boldsymbol{x}$ to obtain $z = \boldsymbol{W}^\top x$, then un project or decode $\boldsymbol{z}$ to get $\hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{z}$ such that $\hat{\boldsymbol{x}}$ is close to $\boldsymbol{x}$ in l_2 distance.

$$\begin{aligned} \text{encode} : \mathbb{R}^D &\to \mathbb{R}^L \\ \boldsymbol{x} &\mapsto \boldsymbol{z} = \boldsymbol{W}^\top \boldsymbol{x} \end{aligned} \qquad \begin{aligned} \text{decode} : \mathbb{R}^L &\to \mathbb{R}^D \\ \boldsymbol{z} &\mapsto \hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{z} \end{aligned}$$

The **reconstruction error** is:

$$\begin{aligned} L(\boldsymbol{W}) &= \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \text{decode}(\text{encode}(\boldsymbol{x}_i))||^2 \\ &== \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i||^2 = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{x}_i||^2 \end{aligned}$$

Solution

We assume the data is centered (mean is zero).
We want to minimize the objective to the constraint that $\boldsymbol{W}$ is an orthogonal matrix.
The optimal solution is

$$\hat{\boldsymbol{W}} = U_L$$

where U_L contains the L eigenvectors with largest eigenvalues of the empirical covariance matrix : $\hat{\Sigma} = \frac{1}{N}\boldsymbol{X}_c^\top \boldsymbol{X}_c$ where $\boldsymbol{X}_c$ is a centered version of the design matrix

It can be shown that minimizing the reconstruction error is equivalent to maximizing the variance of the projected data

PCA in a nutshell

1. Normalize the data
2. Compute the covariance matrix
3. Find the eigenvalue decompositon of the covariance matrix
4. Order the eigenvalues by decreasing order
5. Get the projection matrix whose columns are the L eigenvectors associated with the L largest eigenvalues
6. Project your data using the projection matrix

How are we going to select the number of latent dimensions in PCA ?
Scree plots
Elbow trick
Best reconstruction error on a validation set
imgflip.com

How to choose the number of latent dimensions ?

We cannot use the reconstruction error on a validation set as a way to select the best unsupervised model.
We could do the following.

- Scree plots
- Elbow trick (profile likelihood)

Computational issues

- When $D > N$, it is faster to use the Gram matrix $\boldsymbol{X}\boldsymbol{X}^\top$ instead.
- Computing the SVD decomposition of $\boldsymbol{X}$ is equivalent to computing the eigendecomposition of the covariance matrix but SVD is more numerically stable.
- Consider using randomized SVD for very high dimensional problems.

Beyond PCA

- Variants : Probabilistic PCA, kernel PCA
- Other dimensionality reduction methods : ISOMAP, LLE, MVU, SNE, t-SNE
- Neural networks approaches : auto-encoders, embeddings (word embeddings)