Essy 9: On PCA

Raul Adell

May 2023

1 Introduction

This essay aims to provide help in the mathematical formulation of the method, already explained in previous essays. In order to write this essays formulas given in class are used and also supplementary material from Aluja, T., Morineau, A., Sanchez, G. (2018) Principal Component Analysis for Data Science and Princeton PCA 2003 notes.

2 Overview

rincipal component analysis (PCA) has been called one of the most valuable results from applied linear algebra. PCA is used abundantly in all forms of analysis - from neuroscience to computer graphics - because it is a simple, non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it.

3 Framework: change of basis

The Goal: Principal component analysis computes the most meaningful basis to re-express a noisy, garbled data set. The hope is that this new basis will filter out the noise and reveal hidden dynamics. In general, each data sample is a vector in m- dimensional space, where m is the number of measurement types. Equivalently, every time sample is a vector that lies in an m-dimensional vector space spanned by an orthonormal basis. All measurement vectors in this space are a linear combination of this set of unit length basis vectors.

With this rigor we may now state more precisely what PCA asks: Is there another basis, which is a linear combination of the original basis, that best reexpresses our data set? One might have noticed the addition of the word linear. Indeed, PCA makes one stringent but powerful assumption: linearity. Linearity vastly simplifies the problem by (1) restricting the set of potential bases, and

(2) formalizing the implicit assumption of continuity in a data set. A subtle point it is, but we have already assumed linearity by implicitly stating that the data set even characterizes the dynamics of the system! In other words, we are already relying on the superposition principal of linearity to believe that the data characterizes or provides an ability to interpolate between the individual data points.With this assumption PCA is now limited to reexpressing the data as a linear combination of its basis vectors.

4 Questions remaining

Let **X** and **Y** be m×n matrices related by a linear transformation **PX** is the origin **Y** is a re-representation of that data set. There, **P** is a a rotation and a stretch. If we consider **P** to be a column vector and **X** to be a row, then the previously written scalar product gives as a result **Y**. We recognize that each coefficient of $\mathbf{y_i}$ is a dot product of $\mathbf{x_i}$ with the corresponding row in **PX**.

By assuming linearity the problem reduces to finding the appropriate change of basis. The row vectors $\{p_1, ..., p_m\}$ in this transformation will become the principal components of X. Several questions now arise.

- What is the best way to "re-express" **X**?
- What is a good choice of basis **P**?

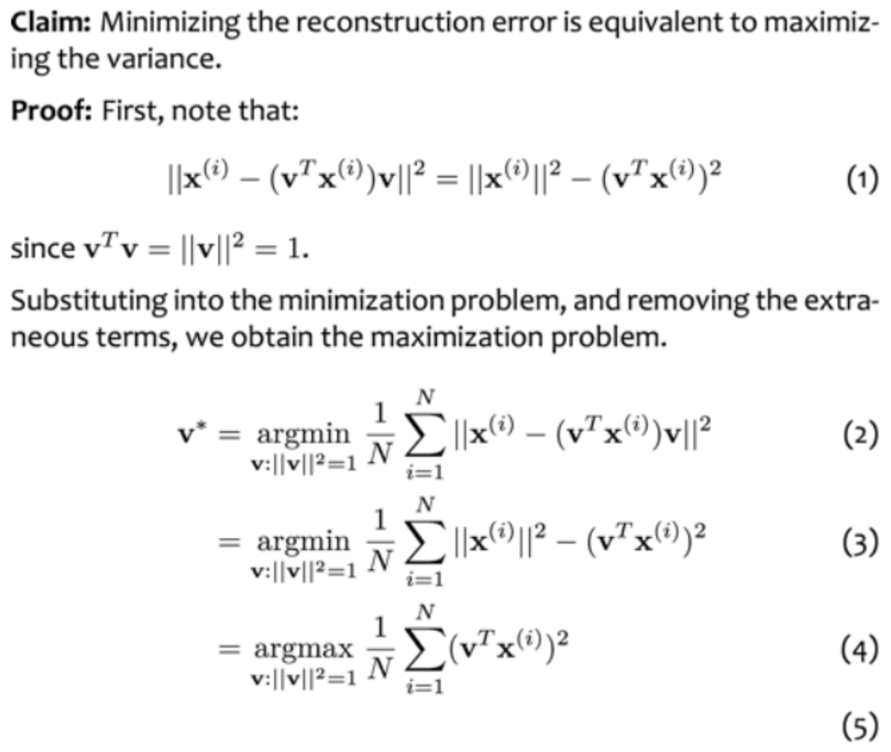
We will see that the answer to these questions goes within the same direction.

5 Maximizing Variance and Minimizing Reconstruction Error

- Maximizing Variance: The first principal component of the dataset corresponds to the direction along which the data exhibits the maximum variance. By projecting the data points onto this principal component, we capture the most significant source of variation. The second principal component is then computed as the direction orthogonal to the first principal component, capturing the second most significant source of variation, and so on. Each subsequent principal component explains as much remaining variance as possible.
- Minimizing Reconstruction Error: To reconstruct the original data points using a lower-dimensional representation, we project the data onto the subspace spanned by the selected principal components. The reconstructed data points are obtained by projecting these projected data points back into the original space.

The following proof is intended to show the equivalence between minimizing construction error and maximizing variance. The way of computing the projection is slightly different. One could interpret as a more geometrical approach.

Under the assumption of centered data and sample variance equal to one, that is standarized data, we can write the following.



Claim: Minimizing the reconstruction error is equivalent to maximizing the variance.

Proof: First, note that:

$$||\mathbf{x}^{(i)} - (\mathbf{v}^T\mathbf{x}^{(i)})\mathbf{v}||^2 = ||\mathbf{x}^{(i)}||^2 - (\mathbf{v}^T\mathbf{x}^{(i)})^2 \tag{1}$$

since $\mathbf{v}^T\mathbf{v} = ||\mathbf{v}||^2 = 1$.

Substituting into the minimization problem, and removing the extraneous terms, we obtain the maximization problem.

$$\begin{aligned} \mathbf{v}^* &= \underset{\mathbf{v}:||\mathbf{v}||^2=1}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}^{(i)} - (\mathbf{v}^T\mathbf{x}^{(i)})\mathbf{v}||^2 && (2)\\ &= \underset{\mathbf{v}:||\mathbf{v}||^2=1}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}^{(i)}||^2 - (\mathbf{v}^T\mathbf{x}^{(i)})^2 && (3)\\ &= \underset{\mathbf{v}:||\mathbf{v}||^2=1}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N} (\mathbf{v}^T\mathbf{x}^{(i)})^2 && (4)\\ & && (5) \end{aligned}$$

While the equivalence has been proven the connection between reconstruction error and (co)variance may seem still unclear. We adress it below:

1. Reconstruction Error: The reconstruction error for a data point $\mathbf{x}_i$ can be defined as the squared Euclidean distance between the original data point and its reconstruction:

$$\text{Reconstruction Error}_i = \|\mathbf{x}_i - \mathbf{x}_{i,\text{reconstructed}}\|^2$$

2. PCA Reconstruction: The reconstruction $\mathbf{x}_{i,\text{reconstructed}}$ is obtained by projecting $\mathbf{x}_i$ onto the subspace spanned by the selected principal components:

$$\mathbf{x}_{i,\text{reconstructed}} = P\mathbf{x}_i = \sum_{j=1}^{k}(\mathbf{x}_i\mathbf{v}_j^T)\mathbf{v}_j$$

3. Variance: The variance of the original data points represents the spread or dispersion of the data around their mean. It can be calculated as the average squared distance from the mean:

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$$

4. Connecting Reconstruction Error and Variance: By substituting the definition of variance into the reconstruction error equation, we have:

$$\text{Reconstruction Error}_i = \|\mathbf{x}_i - \sum_{j=1}^{k}(\mathbf{x}_i \mathbf{v}_j^T)\mathbf{v}_j\|^2 = \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 - \sum_{j=1}^{k}(\mathbf{x}_i \mathbf{v}_j^T)^2$$

   This also shows that the reconstruction error is related to the variance of the data. Minimizing the reconstruction error corresponds to retaining as much of the original variance as possible in the lower-dimensional representation.

6 Covariance and singular value decomposition

In PCA, the goal is to find a lower-dimensional subspace that captures the maximum variance of the data. This can be formulated as the maximization problem:

$$\max_{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k} \text{Var}(X\mathbf{v}_1, X\mathbf{v}_2, \ldots, X\mathbf{v}_k)$$

where X is the centered data matrix, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are the unit-length orthogonal vectors representing the principal components, and k is the desired lower dimension.

Singular Value Decomposition (SVD): SVD is a matrix factorization technique that decomposes a matrix $\mathbf{X}$ into three separate matrices: $\mathbf{X} = \mathbf{UDV^T}$, where:

- $\mathbf{U}$ is an $n \times n$ orthogonal matrix whose columns are the left singular vectors.
- $\mathbf{D}$ is an $n \times d$ diagonal matrix containing the singular values.
- $\mathbf{V^T}$ is the transpose of a $d \times d$ orthogonal matrix whose columns are the right singular vectors.

The singular value decomposition is a generalization of the eigenvalue-eigenvector matrix factorization $\mathbf{A} = \mathbf{EDE^T}$ for non-square matrices. Here $\mathbf{E}$ is built with the normalized eigenvectors of $\mathbf{E}$, which form a basis, while $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{A}$.

1. Given a data matrix X of size $n \times d$ (where n is the number of data points and d is the dimensionality of each data point).

2. Center the data matrix X by subtracting the mean of each feature, resulting in the centered data matrix $\bar{X}$.

3. Compute the covariance matrix Σ of $\bar{X}$, which is given by:

$$\Sigma = \frac{1}{n-1} \bar{X}^T \bar{X}$$

4. Perform an eigendecomposition on the covariance matrix Σ to obtain its eigenvectors and eigenvalues. The eigenvectors represent the principal directions of variation in the data, and the eigenvalues indicate the amount of variance explained by each eigenvector.

5. The eigenvectors of Σ (or their normalized versions) form the columns of the matrix V in the SVD. These are the right singular vectors.

6. The singular values are the square roots of the eigenvalues of Σ, which can be arranged in a diagonal matrix Σ'. The singular values represent the amount of variation explained by each right singular vector.

7. The left singular vectors can be obtained by multiplying the centered data matrix $\bar{X}$ with the matrix V of right singular vectors and scaling by the singular values. Specifically, the left singular vectors are given by:

$$U = \bar{X} V \Sigma'^{-1}$$

   Alternatively, you can normalize the columns of U to have unit length.

7 Bringing everything together

7.1 Connection to Maximizing Variance

The SVD of the centered data matrix X can be used to solve the maximization problem in PCA. The singular values in matrix D represent the square roots of the eigenvalues of the covariance matrix Σ (up to a scaling factor). By selecting the singular vectors corresponding to the largest singular values, we can obtain the principal components that capture the most significant sources of variation in the data. Hence, the maximization problem in PCA is connected to selecting the top k singular vectors of matrix U or V^T with the largest singular values.

7.2 Minimizing Reconstruction Error

In PCA, the goal is also to minimize the reconstruction error, which measures the discrepancy between the original data and its reconstruction using a lower-dimensional representation. This can be formulated as the minimization problem:

$$\min_{\mathbf{x}_{i,\text{reconstructed}}} \|\mathbf{x}_i - \mathbf{x}_{i,\text{reconstructed}}\|^2$$

where $\mathbf{x}_i$ is a data point and $\mathbf{x}_{i,\text{reconstructed}}$ is its reconstruction using a lower-dimensional representation.

7.3 Connection to SVD and Reconstruction Error

The SVD of the centered data matrix X can also be used to solve the minimization problem in PCA. By selecting the k largest singular values and their corresponding singular vectors, we obtain a lower-rank approximation of the original data matrix X as $X_k = U_k D_k V_k^T$. The matrix X_k represents the reconstruction of the data points using the selected principal components. Minimizing the reconstruction error is achieved by selecting the top k singular vectors that correspond to the largest singular values. These singular vectors capture the most important directions of variation in the data and allow us to reconstruct the data points with the least amount of error.

7.4 Recap

Thus, minimizing the reconstruction error is achieved by selecting the k singular vectors that correspond to the largest singular values of the covariance matrix. These singular vectors capture the most important directions of variation and allow us to reconstruct the data points with the least amount of error.
Singular Value Decomposition provides a direct connection between maximizing variance (captured by the singular values) and minimizing reconstruction error (achieved by selecting a subset of singular vectors and singular values). By leveraging the SVD decomposition, PCA allows us to reduce the dimensionality of the data while preserving the most relevant information in terms of variance.

8 Methodology

1. Compute the Covariance Matrix: Given a dataset X with n data points and d features, compute the covariance matrix Σ:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

   where $\mathbf{x}_i$ represents the i-th data point, $\bar{\mathbf{x}}$ is the mean of the dataset, and T denotes the transpose operation.

2. Perform Eigenvalue Decomposition: Perform an eigenvalue decomposition on Σ to obtain the eigenvalues and eigenvectors:

$$\Sigma = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$$

where $\mathbf{V}$ is a matrix containing the eigenvectors and $\mathbf{D}$ is a diagonal matrix containing the eigenvalues.

3. Select Principal Components: Select the first k eigenvectors from $\mathbf{V}$, denoted as $\mathbf{V}_k$. These eigenvectors represent the principal components that capture the most significant sources of variation in the data.

4. Normalize Principal Components: Normalize each column of $\mathbf{V}_k$ to have unit length. This step ensures that the principal components are orthogonal to each other and have comparable magnitudes.

5. Project Data onto Principal Components: Project the original data points onto the subspace spanned by the selected principal components $\mathbf{V}_k$. The projected data points are obtained by multiplying the data matrix X with $\mathbf{V}_k$: $\mathbf{X}' = X \cdot \mathbf{V}_k$

6. Reconstruct Data Points: To reconstruct the data points using the principal components, project the projected data points $\mathbf{X}'$ back into the original space. The reconstructed data points $\mathbf{X}_{\text{reconstructed}}$ are obtained by multiplying $\mathbf{X}'$ with $\mathbf{V}_k^T$: $\mathbf{X}_{\text{reconstructed}} = \mathbf{X}' \cdot \mathbf{V}_k^T$