Vsevolod Shegolev

Updated: 2024-10-18

Principal Component Analysis (PCA)

Intro

- Principal Component Analysis (PCA) is a feature transformation method that converts the original features $\boldsymbol{f}$ into a new set of transformed features $\boldsymbol{p}$, ensuring their linear independence:

$$\boldsymbol{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} \quad \to \quad \boldsymbol{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix},$$

  If the original features are linearly dependent, the data resides in a lower-dimensional space, meaning $m < k$. For clarity, we will assume $m < k$ explicitly.

- The new representation $p_1, ..., p_m$ is constructed as a linear combination of the original features $f_1, ..., f_k$:

$$p_s = \sum_{j=1}^{k} \alpha_{s,j} \cdot f_j,$$

  the coefficients $\alpha_{s,j}$ form the matrix A, which defines the linear transformation from $\boldsymbol{f}$ to $\boldsymbol{p}$.

- The new, usually lower-dimensional, representation $\boldsymbol{p}$ must still be informative. This is achieved by ensuring that $\boldsymbol{p}$ can approximately restore the original features $\boldsymbol{f}$ linearly and with minimal error:

$$\hat{f}_j = \sum_{s=1}^{m} \beta_{j,s} \cdot p_s \approx f_j,$$

  the coefficients $\beta_{j,s}$ form the matrix B, which defines the linear transformation from $\boldsymbol{p}$ back to $\boldsymbol{f}$.

- The objective of PCA is to minimize the reconstruction error $\hat{\boldsymbol{f}} - \boldsymbol{f}$ by finding the optimal linear transformations $A : \boldsymbol{f} \to \boldsymbol{p}$ and $B : \boldsymbol{p} \to \boldsymbol{f}$:

$$R = \sum_{\boldsymbol{x} \in X^\ell} \left\| \hat{\boldsymbol{f}} - \boldsymbol{f} \right\|^2 = \sum_{\boldsymbol{x} \in X^\ell} \| BA\boldsymbol{f} - \boldsymbol{f} \|^2 \to \min_{A,B}.$$

Formalism

Linear Maps

Matrices A (dimension reducer) and B (dimension adder) are linear maps that work oppositely: A reduces the dimension of the original features $\boldsymbol{f}$ to the dimension of the principal components $\boldsymbol{p}$, and B restores, as closely as possible, the original features from the principal components.

$$\boldsymbol{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} \xrightarrow{A} \boldsymbol{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix} \xrightarrow{B} \hat{\boldsymbol{f}} = \begin{pmatrix} \hat{f}_1 \\ \vdots \\ \hat{f}_k \end{pmatrix}$$

This can be written as:

Crumbs on the floor Each data point is represented by three coordinates x, y, z, but z is always 0. Therefore, the data can be represented by just two coordinates:

$$\boldsymbol{f} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \xrightarrow{A} \boldsymbol{p} = \begin{pmatrix} x \\ y \end{pmatrix} \xrightarrow{B} \hat{\boldsymbol{f}} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}.$$

It is straightforward to find the linear transformations A and B:

$$\underbrace{\begin{pmatrix} 1 & 0 & ? \\ 0 & 1 & ? \end{pmatrix}}_{A} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}}_{B} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}$$

NB In the example above:

- The last column of A is arbitrary, so the choice of transformations is not unique.
- A and B are related: $\hat{\boldsymbol{f}} = BA\boldsymbol{f}$, thus $BA = I$.
- Since A and B are non-square, they are non-invertible, so $A = B^{-1}$ does **not** hold.

Crumbs on the table. Now, the third coordinate equals the table height $h = 1$:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \xrightarrow{A} \begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} x \\ y \end{pmatrix} \xrightarrow{B} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

Here, A is the same as before, but no B can restore the original vector exactly.

Formally, if B exists, we could write the system of equations:

$$\begin{pmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \\ \beta_{3,1} & \beta_{3,2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \Rightarrow \begin{cases} 1x + 0y = x \\ 0x + 1y = y \\ \beta_{3,1}x + \beta_{3,2}y = 1 \end{cases}.$$

- The coefficients in the first two equations are determined by the identities $x = x$ and $y = y$.

$$\boldsymbol{p} = A\boldsymbol{f}, \qquad \hat{\boldsymbol{f}} = B\boldsymbol{p}.$$

Matrix Formulation

The feature matrix F and the principal component matrix P are formed by stacking the row vectors $\boldsymbol{f}^\top = (f_1, ..., f_k)$ and $\boldsymbol{p}^\top = (p_1, ..., p_m)$:

$$F := \begin{pmatrix} \boldsymbol{f}_1^\top \\ \vdots \\ \boldsymbol{f}_\ell^\top \end{pmatrix}, \quad P := \begin{pmatrix} \boldsymbol{p}_1^\top \\ \vdots \\ \boldsymbol{p}_\ell^\top \end{pmatrix}$$

In matrix form, the linear maps A and B are applied as follows:

$$P^\top = AF^\top, \qquad \hat{F}^\top = BP^\top,$$

or equivalently, by transposing:

$$P = FA^\top, \quad \hat{F} = PB^\top.$$

Substituting P into $\hat{F}$ yields the following equation:

$$\hat{F} = FA^\top B^\top = F(AB)^\top,$$

The approximation $\hat{F}$ equals F exactly if $AB = I$. Ideally, A would equal B^{-1}, but in general, A and B are non-square and therefore non-invertible.

Pseudoinverse Matrix

$AB = I$ holds if B is the pseudoinverse of A:

$$B = A^+ = \left(A^\top A\right)^{-1} A^\top.$$

$BA = A^+ A = \left(A^\top A\right)^{-1}\left(A^\top A\right) = I$

A^+ is exact if A has full rank, but in general, it does not, so the solution is only approximate:

$$AB \approx I.$$

Geometric Interpretation

Matrices A and B resemble transition matrices between bases:

- A transforms vectors from the original basis of features $f_1, ..., f_k$ into a new space with the basis of principal components $p_1, ..., p_m$. However, since these bases are in different dimensional spaces, this is only an analogy.
- B performs the reverse transformation, converting from the principal component basis back to the original basis (approximately).

Since A and B are related by the pseudoinverse operation and perform inverse transformations, we can focus on one of the matrices. Let it be B.

The basis transition matrix stores the vectors of the new basis in the coordinates of the old basis. As the linear map B transforms principal components into the original features (approximately):

$$\boldsymbol{f} \approx B\boldsymbol{p},$$

it acts similarly to a basis transition matrix from $\boldsymbol{f}$ to $\boldsymbol{p}$, storing the orthogonal basis of principal axes in the coordinates of the original space.

Any basis consists of linearly independent, or orthogonal, vectors, meaning that B stores orthogonal vectors, and $B^\top B = \Lambda$ is diagonal.

Since the choice of B is not unique, we can use this freedom to demand that $B^\top B$ be not just diagonal Λ, but the identity matrix I:

Basis Transition Matrix. If in vector space V, there are two bases: the old one $\mathcal{O} : \boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_n$ and the new one $\mathcal{N} : \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_n$, the vectors of the new basis can be represented as linear combinations of the old basis vectors:

$$\begin{cases} \boldsymbol{\nu}_1 = \alpha_{1,1}\boldsymbol{\omega}_1 + ... + \alpha_{1,n}\boldsymbol{\omega}_n \\ \vdots \\ \boldsymbol{\nu}_n = \alpha_{n,1}\boldsymbol{\omega}_1 + ... + \alpha_{n,n}\boldsymbol{\omega}_n \end{cases}$$

The coefficients $\alpha_{s,j}$ are the coordinates of the new basis vectors in the coordinate system of the old basis. These coefficients form the basis transition matrix (by columns!):

$$A = \begin{pmatrix} \alpha_{1,1} & ... & \alpha_{n,1} \\ \vdots & \ddots & \vdots \\ \alpha_{1,n} & ... & \alpha_{n,n} \end{pmatrix}$$

This matrix transforms coordinates between bases:

$$\{\boldsymbol{\nu}_1\}_{\mathcal{O}} = \begin{pmatrix} \alpha_{1,1} \\ \vdots \\ \alpha_{1,n} \end{pmatrix}_{\mathcal{O}} = A \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}_{\mathcal{N}} = A\{\boldsymbol{\nu}_1\}_{\mathcal{N}}$$

$$\{\boldsymbol{v}\}_{\mathcal{O}} = A\{\boldsymbol{v}\}_{\mathcal{N}}, \quad \{\boldsymbol{v}\}_{\mathcal{N}} = A^{-1}\{\boldsymbol{v}\}_{\mathcal{O}}$$

The choice of matrix B is flexible, allowing us to impose additional constraints. For example, we can require that $B^T B$ be diagonal or even the identity matrix:

$$\exists B : B^\intercal B = I,$$

This implies that B stores not just orthogonal vectors but an **orthonormal** basis of principal components.

Risk Minimization

The objective of PCA is to minimize the restoration error.

In this notation, the empirical risk depends on A and B:

$$\begin{aligned} R &:= \left\|\hat{F} - F\right\|^2 \\ &= \left\|F A^\intercal B^\intercal - F\right\|^2 \to \min_{A,B}. \end{aligned}$$

We can reformulate the objective in terms of the new coordinates P and the transition matrix B by substituting $P = FA^\intercal$, which at least reduces one matrix multiplication:

$$R = \left\|P B^\intercal - F\right\|^2 \to \min_{P,B}.$$

By differentiating R with respect to P and B, we can find the values of P and B at the extremum:

$$\frac{\partial R}{\partial P} = 2(PB^\intercal - F)B = 0$$

$$\Downarrow$$

$$P = FB(B^\intercal B)^{-1}$$

$$\frac{\partial R}{\partial B} = 2P^\intercal (PB^\intercal - F) = 0$$

$$\Downarrow$$

$$B^\intercal = (P^\intercal P)^{-1} P^\intercal F$$

$$\begin{aligned} B &= F^\intercal P\Big((P^\intercal P)^{-1}\Big)^\intercal \\ &= F^\intercal P\Big((P^\intercal P)^\intercal\Big)^{-1} \\ &= F^\intercal P (P^\intercal P)^{-1} \end{aligned}$$

$S = P^\intercal P$ is symmetric, i.e. $S^T = S$

The objective R depends only on the product PB^T, which can result from multiplying any number of different pairs of matrices:

$$PB^\intercal = PIB^\intercal = \underbrace{(P^* R)}_{P}\underbrace{\left(R^{-1} {B^*}^\intercal\right)}_{B^\intercal}$$

We will use the freedom in choosing R and let $P^\intercal P$ and $B^\intercal B$ be diagonal:

- P stores the principal components in their respective coordinates.
- B stores the orthonormal "basis" of principal components in the coordinates of the original space, so $B^\intercal B = I$.

$$\begin{cases} P^\intercal P = \Lambda \\ B^\intercal B = I \end{cases}$$

Earlier, we showed that B could be chosen to store an orthonormal basis, but this wasn't strictly necessary.

It can be demonstrated analytically that it is sufficient to choose R such that $B^T B$ is diagonal, which is enough to ensure $B^T B = I$. This will determine the form of B, which can then be interpreted as a matrix storing an orthonormal basis.

As the proof involves boring linear algebra, we relied on geometric intuition instead (though formal proof is possible!).

Now, we can further simplify the expressions for P and B:

$$\begin{aligned} P &= FB(B^\intercal B)^{-1} = FBI, \\ B &= F^\intercal P (P^\intercal P)^{-1} = F^\intercal P \Lambda^{-1}. \end{aligned}$$

Eliminate P:

$$B\Lambda = F^\intercal F B$$

This means that the columns of B are eigenvectors of $F^\intercal F$:

$$\boldsymbol{b}_j \cdot \lambda_j = (F^\intercal F)\boldsymbol{b}_j.$$

Eliminate B:

$$P\Lambda = FF^{\top}P$$

This means that the columns of P are eigenvectors of $FF^{\top}$:

$$\boldsymbol{p}_j \cdot \lambda_j = (FF^{\top})\boldsymbol{p}_j.$$