

Principal Component Analysis (PCA)

Intro

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

$$\mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} \rightarrow \mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix},$$

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, \dots, p_m is constructed as a linear combination of the original features f_1, \dots, 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 \mathbf{f} to \mathbf{p} .

- The new, usually lower-dimensional, representation \mathbf{p} must still be informative. This is achieved by ensuring that \mathbf{p} can approximately restore the original features \mathbf{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 \mathbf{p} back to \mathbf{f} .

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

$$R = \sum_{\mathbf{x} \in X^\ell} \|\hat{\mathbf{f}} - \mathbf{f}\|^2 = \sum_{\mathbf{x} \in X^\ell} \|BA\mathbf{f} - \mathbf{f}\|^2 \rightarrow \min_{A,B}.$$

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 \mathbf{f} to the dimension of the principal components \mathbf{p} , and B restores, as closely as possible, the original features from the principal components.

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

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:

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

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: $\hat{\mathbf{f}} = BA\mathbf{f}$, 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

This can be written as:

$$\mathbf{p} = A\mathbf{f}, \quad \hat{\mathbf{f}} = B\mathbf{p}.$$

Matrix Formulation

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

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

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

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

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 :

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

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

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, \dots, f_k into a new space with the basis of principal components p_1, \dots, 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):

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

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

Basis Transition Matrix. If in vector space V , there are two bases: the old one $\mathcal{O} : \omega_1, \dots, \omega_n$ and the new one $\mathcal{N} : \nu_1, \dots, \nu_n$, the vectors of the new basis can be represented as linear combinations of the old basis vectors:

$$\begin{cases} \nu_1 = \alpha_{1,1}\omega_1 + \dots + \alpha_{1,n}\omega_n \\ \vdots \\ \nu_n = \alpha_{n,1}\omega_1 + \dots + \alpha_{n,n}\omega_n \end{cases}$$

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!):

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

This matrix transforms coordinates between bases:

$$\{\nu_1\}_{\mathcal{O}} = \begin{pmatrix} \alpha_{1,1} \\ \vdots \\ \alpha_{1,n} \end{pmatrix}_{\mathcal{O}} = A \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}_{\mathcal{N}} = A\{\nu_1\}_{\mathcal{N}}$$

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

The choice of matrix B is flexible, allowing us to impose addi-

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

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

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

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 :

$$\begin{aligned} R &:= \|\hat{F} - F\|^2 \\ &= \|FA^T B^T - F\|^2 \rightarrow \min_{A, B}. \end{aligned}$$

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

$$R = \|PB^T - F\|^2 \rightarrow \min_{P, B}.$$

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

$$\begin{aligned} \frac{\partial R}{\partial P} &= 2(PB^T - F)B = 0 & \frac{\partial R}{\partial B} &= 2P^T(PB^T - F) = 0 \\ \Downarrow & & \Downarrow & \\ P &= FB(B^T B)^{-1} & B^T &= (P^T P)^{-1} P^T F \\ & & B &= F^T P \left((P^T P)^{-1} \right)^T \\ & & &= F^T P \left((P^T P)^T \right)^{-1} \\ & & &= F^T P (P^T P)^{-1} \end{aligned}$$

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

$$PB^T = PIB^T = \underbrace{(P^* R)}_P \underbrace{(R^{-1} B^{*\top})}_{B^T}$$

We will use the freedom in choosing R and let $P^T P$ and $B^T B$ 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 $B^T B = I$.

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

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

$$S = P^T P \text{ is symmetric, i.e. } S^T = S$$

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!).

$$P = FB(B^T B)^{-1} = FBI,$$

$$B = F^T P (P^T P)^{-1} = F^T P \Lambda^{-1}.$$

Eliminate P :

$$B\Lambda = F^T FB$$

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

$$\mathbf{b}_j \cdot \lambda_j = (F^T F) \mathbf{b}_j.$$

Eliminate B :

$$P\Lambda = FF^T P$$

This means that the columns of P are eigenvectors of FF^T :

$$\mathbf{p}_j \cdot \lambda_j = (FF^T) \mathbf{p}_j.$$