Principal components analysis

1 Why change dimensions?

Working with a large number of variables can pose several practical and theoretical problems:

- Complicated visualization: it is impossible to visually represent data beyond 3 dimensions.
- Difficult class separation: in classification problems, the separation between groups may be hidden in a combination of variables rather than in the variables taken individually.
- High computational cost: complex models can become difficult to adjust when the number of variables is large.
- Strong correlations: redundant variables make models unstable or difficult to interpret.

The natural question to ask is therefore: can we reduce the dimension of the dataset without losing too much information?

Reducing the size does not simply mean removing variables. Doing so could result in the loss of information that may be useful to the model. A better approach is to construct new variables, obtained as linear combinations of the initial variables, which summarize the essential information in the dataset. One possible method for doing this is **Principal Component Analysis** (PCA).

2 Principal Component Analysis

PCA is an unsupervised method (without variables to explain) that reduces the dimension of a dataset while retaining as much information as possible. This method is used when there are n observations of p continuous numerical variables with p too "large" to allow for effective modeling or visualization. The method was introduced by Hotelling in Analysis of Complex Statistical Data in 1933.

Common applications

1. Visualization of a multidimensional dataset.
2. Reduction of the number of variables from p to $k \ll p$ to facilitate model construction.
3. Compressing images or signals.
4. Exploring biological, textual, or environmental data.

Examples

1. Compare hockey teams based on six end-of-season statistics.
2. Summarize crime rates among Canadian provinces based on rates for seven different types of crimes.
3. Compress images consisting of 1084×1084 pixels into a few variables.
4. Identify the number of variants of a tumor type based on the expression levels of millions of genes.

2.1 Mathematical formulation

Let $X = \left(X_1, \dots, X_p\right)^\top$ be a random vector composed of p variables, centered and having variance-covariance matrix Σ. Let $\alpha_1 = \left(\alpha_{11}, \dots, \alpha_{1p}\right)^\top$ be a vector of coefficients. We are looking for a linear combination

$$Y_1 = \alpha_1^\top X = \sum_{k=1}^{p} \alpha_{1k} X_k,$$

such that the variance of Y_1 is maximized. The idea is simple: we want to combine p variables into a single one, but while "capturing" as much of the variability as possible.

First, we must add a constraint on α_1, since otherwise we would only have to take $\alpha_{1k} = \pm\infty$ and we would have $\mathrm{Var}(Y_1) = +\infty$, which is definitely maximal! We therefore constrain α_1 so that it has a norm equal to 1.

This amounts to calculating:

$$\max_{\alpha_1^\top \alpha_1 = 1} \mathrm{Var}(Y_1) = \max_{\alpha_1^\top \alpha_1 = 1} \alpha_1^\top \Sigma \alpha_1.$$

This problem is solved using Lagrange multipliers. It leads to the equation

$$\Sigma \alpha_1 = \lambda_1 \alpha_1,$$

where λ_1 is the largest eigenvalue of Σ and α_1 is the associated eigenvector.

This defines the first principal component. The following components are constructed by imposing orthogonality conditions (linear independence) with the previous ones, which amounts to finding the following eigenvectors:

$$\Sigma\alpha_k = \lambda_k\alpha_k, \quad \text{with} \quad \lambda_1 \geq \lambda_2 \geq \dots \lambda_p.$$

The principal components are therefore given by

$$Y_k = \alpha_k^\top X, \quad \text{with } \alpha_k \text{ being the eigenvector associated with } \lambda_k.$$

Note

If $\lambda_1 > \cdots > \lambda_p$, then the principal components are unique, up to sign.

Proof

We seek to calculate

$$\max_{\alpha_1^\top\alpha_1=1} \operatorname{Var}(Y_1) = \max_{\alpha_1^\top\alpha_1=1} \alpha_1^\top\Sigma\alpha_1.$$

The problem is therefore to maximize

$$F(\alpha_1) = \alpha_1^\top\Sigma\alpha_1, \quad \text{s.c.} \quad \alpha_1^\top\alpha_1 = 1.$$

This problem can be rewritten using Lagrange multipliers, i.e., maximize

$$F(\alpha_1, \lambda) = \alpha_1^\top\Sigma\alpha_1 - \lambda(\alpha_1^\top\alpha_1 - 1),$$

where λ is a Lagrange multiplier.
To solve this problem, we differentiate F with respect to α_1 and λ.

$$\begin{cases} \frac{\partial F(\alpha_1,\lambda)}{\partial\alpha_1} = 2\Sigma\alpha_1 - 2\lambda\alpha_1, \\ \frac{\partial F(\alpha_1,\lambda)}{\partial\lambda} = 1 - \alpha_1^\top\alpha_1. \end{cases}$$

Then, we set equal to 0, which gives:

$$\begin{cases} \Sigma\alpha_1 = \lambda\alpha_1 \\ \alpha_1^\top\alpha_1 = 1 \end{cases}.$$

The second equation is, of course, our constraint. The first equation is the one we are interested in. Using this equation and the definition of eigenelements, we deduce that

1. α_1 is an eigenvector (normed) of Σ;
2. λ is the corresponding eigenvalue.

We therefore have that

$$\mathrm{Var}(Y_1) = \alpha_1^\top \Sigma \alpha_1 = \lambda \alpha_1^\top \alpha_1 = \lambda.$$

Since we want to maximize this quantity, we conclude that:

1. $\lambda = \lambda_1$, the largest eigenvalue of Σ;
2. α_1, the corresponding normalized eigenvector.

We then continue with the calculation of the second component. Here, we pursue two objectives simultaneously:

1. Preserve the maximum variation present in X;
2. Simplify the dependency structure to facilitate interpretation and ensure the numerical stability of any methods that will use the principal components obtained.

Given Y_1, the second principal component $Y_2 = \alpha_2^\top X$ is defined such that

1. $\mathrm{Var}(Y_2) = \alpha_2^\top \Sigma \alpha_2$ is maximized;
2. $\alpha_2^\top \alpha_2 = 1$;
3. $\mathrm{Cov}(Y_1, Y_2) = 0$.

However, we have that

$$\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(\alpha_1^\top X, \alpha_2^\top X) = \alpha_1^\top \Sigma \alpha_2 = \alpha_2^\top \Sigma \alpha_1 = \lambda_1 \alpha_2^\top \alpha_1.$$

We are therefore looking for the vector α_2 that maximizes:

$$F(\alpha_2, \lambda, \kappa) = \alpha_2^\top \Sigma \alpha_2 - \lambda(\alpha_2^\top \alpha_2 - 1) - \kappa(\alpha_2^\top \alpha_1 - 0).$$

As with the first component, we differentiate F with respect to α_2, λ, and κ.

$$\begin{cases} \frac{\partial F(\alpha_2,\lambda,\kappa)}{\partial \alpha_2} = 2\Sigma\alpha_2 - 2\lambda\alpha_2 - \kappa\alpha_1 \\ \frac{\partial F(\alpha_2,\lambda,\kappa)}{\partial \lambda} = 1 - \alpha_2^\top \alpha_2 \\ \frac{\partial F(\alpha_2,\lambda,\kappa)}{\partial \kappa} = -\alpha_2^\top \alpha_1 \end{cases}$$

By setting the equations equal to 0, we find the two constraint equations, as well as

$$2\Sigma\alpha_2 - 2\lambda\alpha_2 - \kappa\alpha_1 = 0.$$

Multiplying this equation on the left and right by $\alpha_1^\top$, we find

$$2\alpha_1^\top \Sigma \alpha_2 - 2\alpha_1^\top \lambda \alpha_2 - \kappa \alpha_1^\top \alpha_1 = 0.$$

Now, $\alpha_1^\top \Sigma = \lambda_1 \alpha_1^\top$, and $\alpha_1^\top \alpha_1 = 1$, so

$$2\alpha_1^\top \lambda \alpha_2 - 2\alpha_1^\top \lambda \alpha_2 - \kappa \alpha_1^\top \alpha_1 = 0 \implies -\kappa = 0.$$

Substituting this result, we obtain

$$\Sigma \alpha_2 = \lambda \alpha_2.$$

and therefore λ is another eigenvalue of Σ. Since

$$\mathrm{Var}(Y_2) = \alpha_2^\top \Sigma \alpha_2 = \alpha_2^\top \lambda \alpha_2 = \lambda,$$

we have that this variance is maximal if $\lambda = \lambda_2$, the second largest eigenvalue of Σ, and consequently α_2 is the corresponding normalized eigenvector.
We can generalize this result using successive maximizations. We conclude that

$$Y_k = \alpha_k^\top X,$$

where α_k is the normalized eigenvector associated with λ_k, the kth largest eigenvalue of Σ.

It is possible to have a more compact representation of PCA using matrices. Let $A = (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^{p \times p}$ be the matrix whose columns are the eigenvectors. We have $Y = AX$ and the covariance of the principal components is written as

$$\mathrm{Var}(Y) = A^\top \Sigma A.$$

Properties of A

1. $A^\top A = AA^\top = I_p$;
2. $A^\top = A^{-1}$;
3. $\Sigma A = A\Lambda$, where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$;
4. $\mathrm{Var}(Y) = A^\top \Sigma A = \Lambda \implies \mathrm{Cov}(Y_k, Y_l) = 0$ if $k \neq l$ and $\mathrm{Var}(Y_k) = \lambda_k \geq \mathrm{Var}(Y_l) = \lambda_l$ if and only if $k \geq l$.

Proofs

1. By construction, the columns of A are pairwise orthogonal and have norm 1 (see constraints on eigenvectors). The matrix A is therefore an orthogonal matrix. And so $A^\top A = AA^\top = I_p$.
2. Similarly, since A is orthogonal, we have $A^\top = A^{-1}$.

3. The result is immediate by multiplying the matrices.

4. We have $\mathrm{Var}(Y) = A^\top \Sigma A = A^\top A \Lambda = \Lambda$. Since Λ is a diagonal matrix, $\mathrm{Cov}(Y_k, Y_l) = 0$ if $k \neq l$ (because it is not on the diagonal) and since $\lambda_1 \geq \cdots \geq \lambda_p$, we have $\mathrm{Var}(Y_1) \geq \cdots \geq \mathrm{Var}(Y_p)$.

An overall measure of the variation present in the data is given by the trace of the matrix Σ:

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i = \sum_{k=1}^{p} \mathrm{Var}(Y_k).$$

The proportion of variation explained by the principal component Y_k is therefore given by the ratio between the eigenvalue k and the sum of the eigenvalues:

$$\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p} = \frac{\mathrm{Var}(Y_k)}{\mathrm{tr}(\Sigma)}.$$

Similarly, the first m components explain

$$100\% \times \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{k=1}^{p} \lambda_k} = 100\% \times \frac{\sum_{k=1}^{m} \mathrm{Var}(Y_k)}{\sum_{k=1}^{p} \mathrm{Var}(Y_k)}$$

of the variability in the variables.

3 Practice of PCA

3.1 Estimation of the variance-covariance matrix

In practice, the variance-covariance matrix Σ is unknown. To perform PCA, it is necessary to estimate Σ from a random sample $X_1, \ldots, X_n$ of independent realizations of X. An (unbiased) estimator of Σ is given by

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(X_i - \widehat{X}\right)^\top,$$

where $\overline{X}$ is the empirical mean of the sample.

The matrix $\widehat{\Sigma}$ thus obtained is symmetric with real coefficients and therefore diagonalizable. It admits a spectral decomposition of the form

$$\widehat{\Sigma} = \widehat{A}\widehat{\Lambda}\widehat{A}^\top,$$

where $\widehat{A}$ is an orthogonal matrix whose columns are the estimators of the eigenvectors of Σ and $\widehat{\Lambda}$ is a diagonal matrix containing the estimators of the eigenvalues of Σ, assumed to be ordered in descending order.

The principal components are obtained by projecting the observations X_i onto the basis of eigenvectors:

$$Y_i = \widehat{A}^\top X_i.$$

3.2 Some remarks

Sensitivity to the scale of $X_1, \dots, X_p$

Since we are looking for a linear combination of $X_1, \dots, X_p$ that maximizes the variance, a variable X_k with a large variance will have a disproportionate weight in the principal components, which can distort the interpretation. One example is distance measurement. Expressing distance in meters rather than kilometers would multiply the variance of this variable by 1 million $(10^3)^2$. This variable would therefore have a major weight in all components.
Thus, if the variables are expressed in different units or have very different orders of magnitude, it is recommended to standardize the variables before performing PCA. This is equivalent to performing PCA on the correlation matrix.

First step in predictive analysis

Sometimes PCA is performed because we want to predict the value of variable Z from the values of variables $X_1, \dots, X_p$, but p is simply too large. In this case, PCA is applied to obtain the first $k \ll p$ principal components $Y_1, \dots, Y_k$ and these are used to predict Z. Since the principal components retain most of the information contained in the original variables, this is generally a reasonable approach.

Axis rotation and representation quality

Since the matrix A is orthogonal, the matrix product $Y = A^\top X$ represents a rotation of the variable space. The new axes correspond to the successive orthogonal directions of maximum variation, assuming that $\lambda_1 > \dots > \lambda_p$. Thus, $Y_i = A^\top X_i$ gives the coordinates of the observation X_i in the new axis system. We call Y_{ik} the score of the observation X_i on the principal axis k and calculate it as

$$Y_{ik} = \alpha_k^\top X_i = \sum_{l=1}^{p} \alpha_{kl} X_{il}.$$

The quality of the representation of observation i on axis k is given by

$$Q_{ik} = \frac{Y_{ik}^2}{d^2(0, Y_i)} = \frac{Y_{ik}^2}{\sqrt{Y_{i1}^2 + \cdots + Y_{ip}^2}}.$$

3.3 Choosing the number of components

A key issue in PCA is choosing how many principal components to retain. Retaining too many does not reduce the dimension, and retaining too few can result in the loss of relevant information. Here are the main rules of thumb used:

1. **80% rule**: Retain the minimum number of components necessary to explain at least 80% of the total variance. This threshold is arbitrary, but it often provides a good intuition.
2. **Kaiser's rule**: If PCA is performed using the correlation matrix, then the average eigenvalue is 1. It is recommended to keep only those components with an eigenvalue greater than the average eigenvalue, i.e., 1.
3. **Joliffe's rule**: A stricter variant of Kaiser's rule, which suggests keeping components with an eigenvalue greater than 0.7 for PCA performed using the correlation matrix.
4. **Cattell's rule (or elbow rule)**: Plot the eigenvalues λ_k as a function of their rank k and look for a breakpoint in the decline. Beyond this point, additional components explain little additional variance.

These rules are decision-making tools, but the choice of the number of components also depends on the context, the objectives of the analysis, and ease of interpretation.