Principal Component Analysis (PCA)

panko.aliaksandr

September 2020

Contents

1 Introduction

Let's explain PCA on a real world example.

- Banks are issuing loans, making money if a borrower returns the loan with the interest back.
- The problem is to define whom to give a loan to receive the money back.
- There is much information about a client available:
  - Salary
  - Address
  - Sex
  - Marital status
  - Credit Historical Score
  - House Value
  - Car Price
  - Number of children etc.

- Let's name each piece of information a factor. It is clear that some factors (say, Salary) are more important than others (say, Sex) to define if a person returns the loan back.
- The more relevant information a factor gives the more important for the decision making process this factor is.
- Moreover, some factors provide overlapping information. Expensive car, expensive house or high salary - all this says that a person is probably rich.
- As soon as we get salary information, we can conclude something about the house and the car. So, the car and house information becomes redundant.
- Having more information is always better, but for modeling as smaller the number of factors as better. This makes models more reliable, understandable and fast.
- This leads us to the next question: "Can we somehow aggregate initial information to create new more representative factors?"
- In our example, we need something like: "Wealthiness", "Health", "Credit History", "Social Status".
- Wealthiness in this case includes combination of salary, house, car and other assets. Social Status includes marital status, number of kids, race, age, address etc.
- Now it is clear that newly created factors (aggregated from initial ones) give us more or less the same information, but the amount of the factors are less.
- The purpose of PCA is to aggregate initial factors into new factors (linear combination of original ones) in a way to avoid information overlap.
- If we additionally could order such factors by the quantity/significance of information provided, we can just drop the least significant information (factors with the smallest amount of new relevant information added). This would result in smaller number of factors that is much better for modeling.

Resume

- PCA is a technique which creates new more informative factors as a linear combination of the original factors in a way that each new factor adds only new information.
- Ex. If the first factor represent money related features, the second could aggregate health related features (assuming that there is no connection between money and health).

2 Steps of PCA

Let's go through all the steps to deeply understand how it actually works.

The first idea that comes to mind is correlation analysis. If some factors are very highly correlated there is a good idea to eliminate one of them isn't it?

2.1 Variance and Information

The first core idea to understand is how a factor variance is connected with the information the factor brings.

- The higher factor variance is the more information factor brings.
- To get it, let's consider couple of examples.
  - Assume that all your customers are male. In this case factor Sex has no variation at all, so the variance is 0 and based on this factor we cannot improve our expectation about the borrower quality.
  - Let's assume that factor Salary has 3 values:
    - poor(1)
    - mid(2)
    - rich(3)

    factor Health has also 3 values:
    - bad(1)
    - mid(2)
    - good(3)

    Assume further that they are evenly distributed among the customers. We can create a $Scorefactor = 2 * Salary + 1 * Health$ (2 is just example, because I think that salary is more important). Now Score factor has values from 3 to 9 and, since the values are distributed evenly, higher variance.
  - The increase in variance clearly comes with increase of information: score 9 gives you more information than just good health.
  - One more important point. Any factor has the same amount of points(observations), so the factor with the biggest variance helps to distinguish points better, increasing the distance between them. So the points become not so similar.

So, we need to construct factors with biggest possible variance. It is important to understand that, when working with variances, data must be standardized to prevent a scale bias. Let me explain:

- variance is a measure of spread, so if the spread is big, the variance is also big.

- if the mean is 1000 with the spread ±5, intuitively there is almost no deviation from the mean.
- if the mean is 0.1 and the deviation is ±0.05 (which is btw 50%) it is obvious that the spread is huge. But in terms of just single number in the first case the deviation will be bigger, which is misleading.

Resume

- All original factors must be standardized to represent variance comparably
- The factor with bigger variance contains more information.

2.2 Original Data Standardization

To standardize a factor, we need to calculate factor values mean and standard deviation. Then apply the next formula to each factor observation.

$$data_{new} = \frac{data_{old} - mean}{\sigma}$$

This makes the factors values comparable, since after this step all the factors has values with mean = 0 and standard deviation = 1.

2.3 Covariance Matrix Calculation

This step consists of 2 substeps:

- Put each factor's values as columns into a matrix.
- Compute corresponding covariance matrix.

This step analyses how correlated original factors are. The values of the covariance matrix show which factor provide overlapping information.

2.4 Wanted Factors Properties

This is the most difficult part. Now we need to find the set of factors which have the biggest possible variance and "do not overlap". "Do not overlap" mean "do not contain redundant information" and mathematically speaking are uncorrelated.

- Let's assume that initially all factors are linearly independent, otherwise we can logically exclude them.
- Ex. house price and salary are independent (just probably highly correlated), when Income, Savings, Spendings are definitely not independent, since Savings = Income - Spendings

- Principle components (new factors) must be mutually independent and uncorrelated to efficiently explain the variance/difference (provide non-overlapping information) of original data.
- The most natural idea to create a set of linearly independent and uncorrelated vectors is to create an orthogonal basis in the original factor space.
- This can be achieved by Gram–Schmidt orthogonalization process. However, the initial vector is required. Having it all others are constructed automatically.
- We want to find a vector with norm = 1, otherwise, one can get infinitely big variance.

Resume

- Wanted factors are set of orthogonal vectors (linearly independent, uncorrelated) which are built as a linear combination of initial factors.
- Wanted factors contain maximum of available information, maximizing the factors variance.
- Wanted factors must be normalized.

2.5 Optimization Problem

Let's connect this problem to portfolio management problem.

- We need to organize our original factors in a way to get a new factor with the maximum possible variance.
- We are supposed to use linear combination, so as soon as we know correct weights, the construction can be easily done.
- This task is absolutely the same as "construct a maximum variance portfolio" with the same constraints, since both problems search for the same weights (in portfolio management factors are just assets' time-series)
- We know that portfolio variance is $w\Sigma w$, where w is vector with assets weights in the portfolio.
- So, to maximize FPC variance we need to maximize "portfolio" variance: $w\Sigma w \to Max$
- This is optimization problem which can be solved via Lagrange multipliers. Having the constraint $\|\mathbf{w}\| = \mathbf{w}^\top \mathbf{w} = 1$ we get: $\mathbf{w}^\top \mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$
- differentiating, we obtain $\mathbf{C}\mathbf{w} - \lambda\mathbf{w} = 0$, which is the eigenvector equation, since $Cw = \lambda w$. This means that some eigenvector indeed maximizes variance.

- Let's prove that we need the one with maximal eigenvalue. To do this let's substitute any eigenvector (accounting for the constraint) into the objective function, which gives $\mathbf{w}^\top \mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1) = \mathbf{w}^\top \mathbf{C}\mathbf{w} = \lambda \mathbf{w}^\top \mathbf{w} = \lambda$. Since the objective function should be maximized, λ must be the largest eigenvalue.

Resume

- Now we know that Eigenvector with the biggest eigenvalue contains weights which should be applied to original factors to get the First Principal Component with maximum possible variance. No other combination of the original factors gives a vector with bigger variance
- The 2-nd, 3-rd etc. eigenvectors are used to maximize remaining variance since they are orthogonal to the first one, so uncorrelated, hence, do not contain overlapping information.
- It sounds quite intuitive that to be uncorrelated vectors must be orthogonal. BTW in statistical area dot product is frequently defined as Cov(x,y), so $< x, y >= 0$ means orthogonal and at the same time uncorrelated. But the main idea is that such choice helps to avoid information overlap.
- Important to pick eigenvectors in descending order of the corresponding eigenvalues! As was proved before it provides highest variance.

2.6 Feature Vector Construction and original data transformation

Now we have required number of the vectors with the highest variance. All we need now is to get the final principal components, which are linear combination of the original factors with the corresponding weights. The weights are just the coordinates of found vectors.

$$pc_i = Mw_i$$

$$PrincipalComponents = OriginalStandardizedFactors * FeatureVector$$

The last step is the dimension reduction: one can include in Feature Vector only first several vectors, reducing dimension of the original factors space/matrix.

2.7 Referenses

Check the references! They contain some great visualization and additional explanations

1. post1
2. post2
3. post3