10.1 What's PCA?

Principal Components Analyses (PCA) is a commonly used method to transform the original data into a set of linearly independent representations of each dimension through linear transformation. It can be used to extract the main feature components of data, in other words, **it is often used to reduce the dimension of high-dimensional data**.

10.2 Effect of Dimensional Reduction

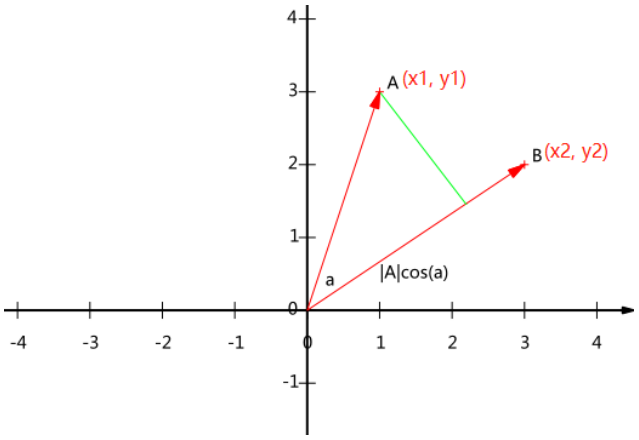
Dimensionality reduction means the loss of information, but the given data is often having frequent relevance itself. We should try to minimize the loss of information while reducing dimensionality. For example, from the data of Amazon store, we can know from experience that "page views" and "visitors" tend to have a strong correlation, which means that if we delete one of the indicators of page views or visitors, we will not lose too much information. Therefore, we can remove one to reduce the complexity of the machine learning algorithm.

- Data is easier to process and use in low dimension.
- Important features can be clearly displayed in the data.
- If there are only two or three dimensions, it is easier to visualize.
- Remove data noise.
- Reduce algorithm overhead.

10.3 Basic Concepts of PCA

10.3.1 Vector Representation

Suppose that A and B are two-dimensional vectors A = (x1, y1) = (3, 1), B = (x2, y2) = (3, 2). On the 2D plane, A and B can be represented by two directed line segments from the origin, as shown in the following figure:



Now let's draw A vertical line from A to B, the intersection of the perpendicular line and B is called the projection of A onto B, and let's say that the angle between A and B is α. We express the inner product in a familiar form:

$$A \cdot B = |A| \cdot |B| \cdot \cos(\alpha) \tag{1}$$

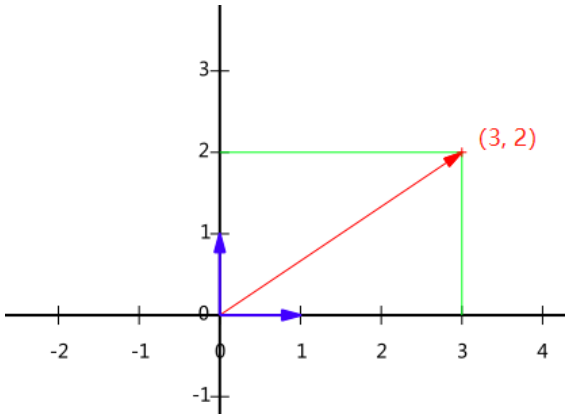
The inner product of A with B is equal to the length of the projection from A to B times the magnitude of B. If the modulus of B is assumed to be 1, that is, |B|=1, then it becomes:

$$A \cdot B = |A| \cos(\alpha) \tag{2}$$

In other words, if the magnitude of vector B is 1, then the inner product of A and B is equal to the vector length of the projection of A onto B.

10.3.2 Bases of Vector

The vectors (x, y) can be represented as linear combinations, which is **$x \cdot (1, 0)^T + y \cdot (0, 1)^T$**. (1,0) and (0,1) here are called a base in 2D. General requirements for base are the length is 1 and orthogonal to each other. For example, the (1, 0) and (0, 1) in the figure below.

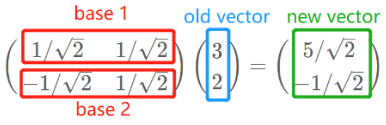


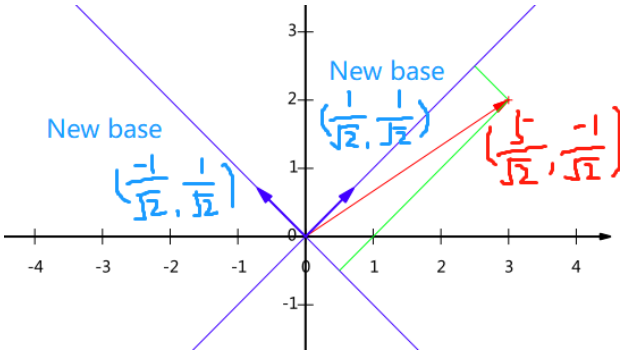
A = (x1, y1) = (3, 1) = $3 \cdot (1, 0)^T + 1 \cdot (0, 1)^T$; B = (x2, y2) = (3, 2) = $3 \cdot (1, 0)^T + 2 \cdot (0, 1)^T$

According to the geometric meaning of the inner product, the coordinates of the dot product with the basis on the new basis can be obtained. In fact, for any vector you can always find a vector whose magnitude is 1 in the same direction, just divide the two components by the magnitude. The following figure is a schematic diagram of the new base and the coordinate value (3, 2) on the new base:

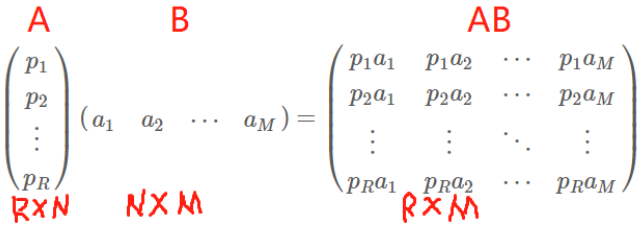
10.3.3 Base Transformation

If we think about the example above, we want to make the transformation of (3, 2) to the coordinates on the new basis. We take the inner product of (3, 2) with the first base as the component of the first new coordinate, and then we take the inner product of (3, 2) with the second base as the component of the second new coordinate. In fact, we can express this transformation simply in terms of matrix multiplication:





In general, if we have M n-dimensional vectors in matrix B, now we want to transform it with R new bases of n dimensions in matrix A. The mth column in matrix AB is the result of mth vector in matrix B after transformation. The mathematical representation is:



- pi in matrix A is a row vector, representing the ith basis.
- aj in matrix B is a column vector, representing the jth raw data.
- the mth column in matrix AB is the output of am after transformation.

10.4 Covariance matrix and optimization

We discussed above that choosing different bases can give different representations to the same set of data, and if the number of bases is less than the dimension of the vector itself, the dimensionality can be reduced. But there is still a crucial question: how to choose the best bases? In other words, if I have a set of n-dimensional vectors, and now I want to reduce it to K dimensions, how do I choose K bases in order to keep the information as much as possible?

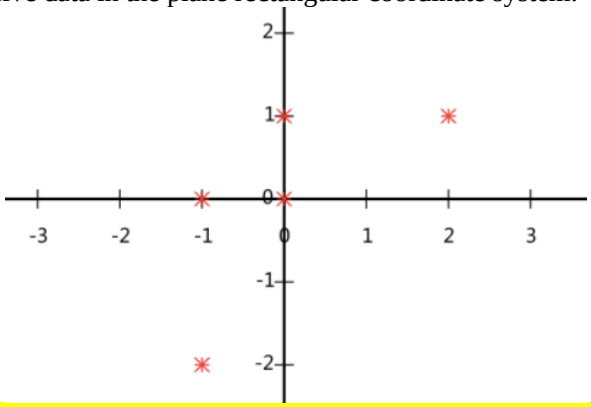
To avoid too abstract a discussion, let's continue with a concrete example. Assume that the data consists of five records and represent them in matrix form:

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

Each of these columns contains one data record. For the convenience of subsequent processing, the equalization should be done first. It is easy to know the mean of the first field is 2, the mean of the second field is 3, so after the transformation:

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

We can look at the five data in the plane rectangular coordinate system:



How do we choose these bases to keep as much original information as possible? If the projection is to the X-axis, the leftmost two points will overlap, and the middle two points will overlap. Likewise, if we project on the Y-axis, the top two points and the two points on the X-axis will overlap, which is a huge loss of information, and neither the X-axis nor the Y-axis is the best projection choice. We can visually see that if we project onto a diagonal line that goes through the first and third quadrants, we can still distinguish the five points after the projection. **Therefore, the principle is we want such a projection so that points can be diffused as much as possible.**

10.4.1 Variance

We want the projection to make points be as dispersed as possible, and this dispersion can be expressed by the mathematical variance. Here, the variance of a field can be regarded as the mean of the sum of squares of the difference between each element and the field mean, that is:

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m}(a_i - \mu)^2$$

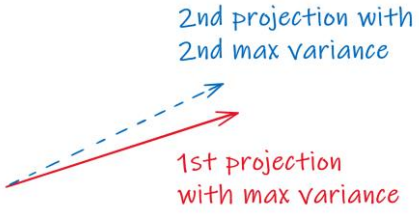
The mean of each field above has been changed to 0, so the variance can be expressed as the sum of squares of each element divided by the number of elements:

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m}a_i^2$$

Therefore, the above problem is expressed as: find a base so that all data is transformed into the coordinate representation on this base, the variance value is the largest.

10.4.2 Covariance

For the problem where the two dimensions go down to one dimension, just find the direction that maximizes the variance. **But for higher dimensions, we first find a direction that maximizes the variance of the projection, so that we can choose the first direction, and then we need to choose the second projection direction.** The second projection direction must be very close to the first projection direction which is shown as below.



In the above situation, although these two projections have as maximum variance as possible, but they do not satisfy the requirements for bases. They are not orthogonal. Therefore, we should make the bases orthogonal, which can be represented as correlation of them is 0. It means that these two bases are absolutely independent so that the two fields are completely independent. Mathematically, the covariance of two fields can be used to represent their correlation.

$$cov(A,B) = E[(X - EX) \cdot (Y - EY)] \tag{3}$$

Since the mean value of each field has been set to 0, then:

$$Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m} a_i b_i$$

If we reduce a set of N-dimensional vector to K dimensional vector (0 < K < N), now we get the optimization goal is:

- to choose K unit orthogonal bases (covariance = 0).
- to find the field of variance is as large as possible.

10.4.3 Covariance Matrix

The optimization goal is derived above, and the ultimate goal is closely related to the in-field variance and the inter-field covariance. Therefore, we introduce covariance matrix which can represent them at the same time. Suppose there are only two fields of a and b, then we will form a matrix X by rows:

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

Then we get **covariance matrix by taking X times the transpose of X, times the coefficient 1/m:**

$$\frac{1}{m}XX^{\top} = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix}$$

We can easily figure out covariance matrix is a symmetric matrix, whose main diagonal lines, on the condition of mean of each field is 0, are the variances of each field respectively, while other elements indicate the covariance of the two fields.

Based on the optimization goal above, we should make variance as much as possible and make sure the covariance equals 0. **In other words, we should make the elements in main diagonal lines of covariance matrix as much as possible and make sure other elements of covariance matrix are 0. This is our goal at the moment. It can be achieve by diagonalizing the covariance matrix.**

10.4.4 diagonalizing the covariance matrix

As we know above, the covariance matrix C is a symmetric matrix. In linear algebra, the real symmetric matrix has a series of very good properties:

- the eigenvectors corresponding to different eigenvalues of the real symmetric matrix must be orthogonal.
- set the eigenvalue lambda multiplicity as r, then there must be r linearly independent eigenvectors corresponding to lambda, so r eigenvectors can be normalized.

From the above two, it can be seen that a real symmetric matrix with n rows and n columns must find n unit orthogonal eigenvectors. Let the n eigenvectors be e1, e2, ... and en.

$$E = \begin{pmatrix} e_1 & e_2 & \cdots & e_n \end{pmatrix}$$

$$E^{\mathsf{T}} C E = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

- Λ: diagonal matrix, the diagonal elements for the corresponding eigenvector of the eigenvalue
- P = E^T is the matrix arranged in rows after the eigenvectors of the covariance matrix are unitized, in which each row is an eigenvector of C. If the set P according to Λ characteristic value from big to small, will feature vector are arranged from top to bottom, with P matrix multiplied by the original data of K line before matrix X, get the need after the dimension reduction of data matrix Y, Y = PX.

10.5 Algorithm Process for PCA

Suppose we have m pieces of n-dimensional data:

1) Divide original data into a matrix X with n rows and m columns
2) Zero averaging of each row of X, i.e. subtracting the mean of this row
3) Find the covariance matrix C = (XX^T) / m
4) Find the eigenvalues of the covariance matrix and the corresponding eigenvectors
5) Arrange the eigenvectors into a matrix in rows from top to bottom according to the corresponding eigenvalues, and take the first k rows to form a matrix P
6) Y = PX is the data after dimension reduction to k

10.6 Example for PCA

Question: reduce the following two-dimensional data to one dimension by PCA:

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

Answer:

1) Divide original data into a matrix X with n rows and m columns

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

2) Zero averaging of each row of X, i.e. subtracting the mean of this row

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

3) Find the covariance matrix C = (XX^T) / m

$$C = \frac{1}{5}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

4) Find the eigenvalues of the covariance matrix and the corresponding eigenvectors
Eigenvalues:

$$\lambda_1 = 2, \lambda_2 = 2/5$$

Eigenvectors:

$$c_1\begin{pmatrix} 1 \\ 1 \end{pmatrix}, c_2\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Normalizations:

$$\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

5) Arrange the eigenvectors into a matrix in rows from top to bottom according to the corresponding eigenvalues, and take the first k rows to form a matrix P

$$P = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Verify the diagonalization of the covariance matrix C:

$$PCP^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}\begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix}\begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix}$$

6) Y = PX is the data after dimension reduction to k

$$Y = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

Effect show after dimension reduction: