Computer vision for dummies

Go to...

A geometric interpretation of the covariance matrix

Contents [hide] [hide]

Introduction

In this article, we provide an intuitive, geometric interpretation of the covariance matrix, by exploring the relation between linear transformations and the resulting data covariance. Most textbooks explain the shape of data based on the concept of covariance matrices. Instead, we take a backwards approach and explain the concept of covariance matrices based on the shape of data.

In a previous article, we discussed the concept of variance, and provided a derivation and proof of the well known formula to estimate the sample variance. Figure 1 was used in this article to show that the standard deviation, as the square root of the variance, provides a measure of how much the data is spread across the feature space.

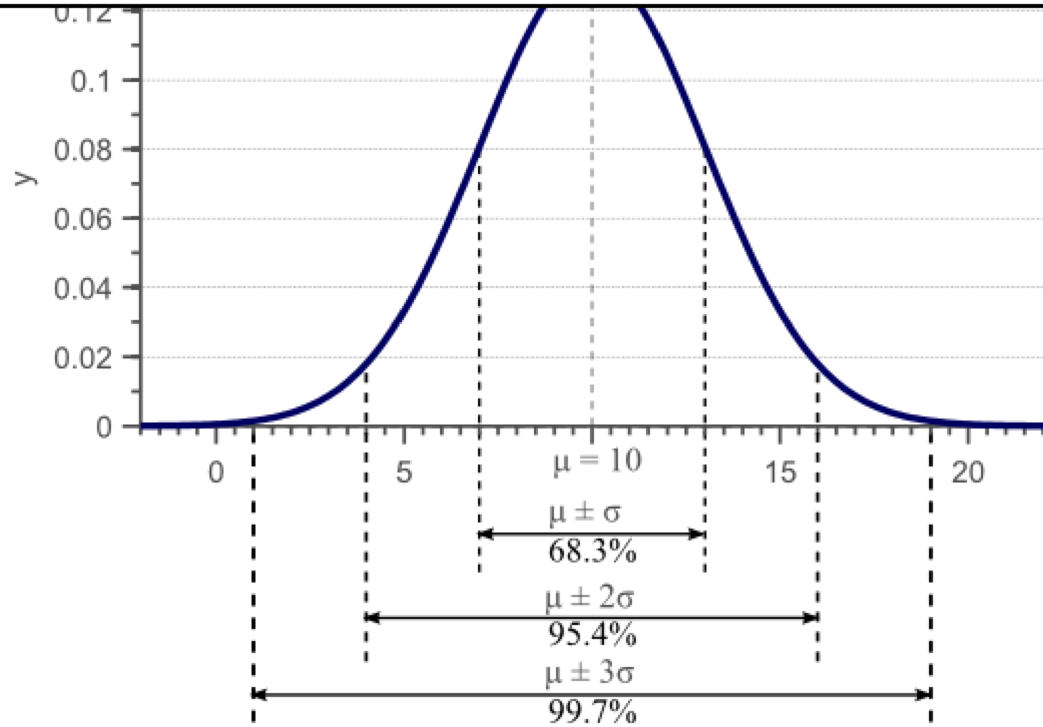


Figure 1. Gaussian density function. For normally distributed data, 68% of the samples fall within the interval defined by the mean plus and minus the standard deviation.

We showed that an unbiased estimator of the sample variance can be obtained by:

$$\begin{aligned}\sigma_x^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2 \\ &= \mathbb{E}[(x-\mathbb{E}(x))(x-\mathbb{E}(x))] \\ &= \sigma(x,x)\end{aligned} \tag{1}$$

However, variance can only be used to explain the spread of the data in the directions parallel to the axes of the feature space. Consider the 2D feature space shown by figure 2:

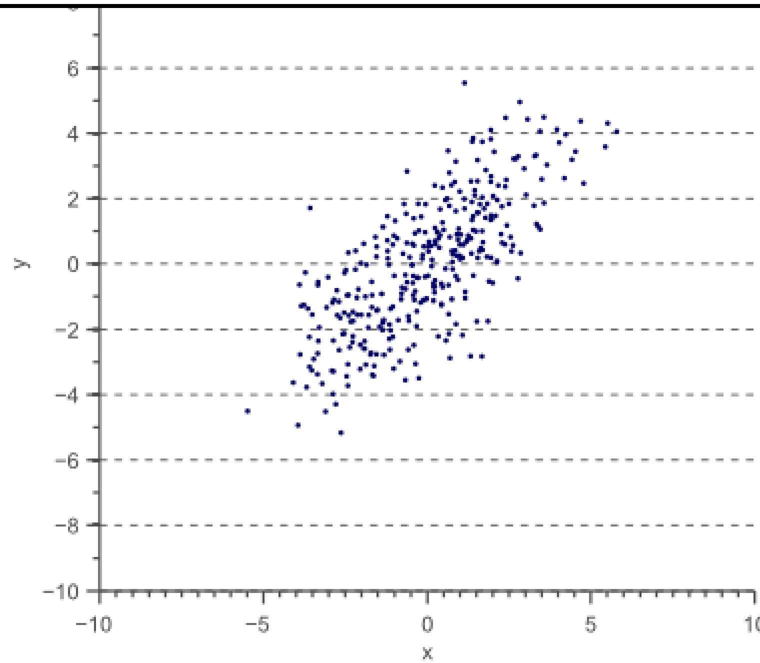


Figure 2. The diagnoal spread of the data is captured by the covariance.

For this data, we could calculate the variance $\sigma(x,x)$ in the x-direction and the variance $\sigma(y,y)$ in the y-direction. However, the horizontal spread and the vertical spread of the data does not explain the clear diagonal correlation. Figure 2 clearly shows that on average, if the x-value of a data point increases, then also the y-value increases, resulting in a positive correlation. This correlation can be captured by extending the notion of variance to what is called the 'covariance' of the data:

$$\sigma(x,y) = \mathbb{E}[(x - \mathbb{E}(x))(y - \mathbb{E}(y))] \tag{2}$$

For 2D data, we thus obtain $\sigma(x,x)$, $\sigma(y,y)$, $\sigma(x,y)$ and $\sigma(y,x)$. These four values can be summarized in a matrix, called the covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix} \tag{3}$$

If x is positively correlated with y, y is also positively correlated with x. In other words, we can state that $\sigma(x,y) = \sigma(y,x)$. Therefore, the covariance matrix is always a symmetric matrix with the variances on its diagonal and the covariances off-diagonal. Two-dimensional normally distributed data is explained completely by its mean and its 2×2 covariance matrix. Similarly, a 3×3 covariance matrix is used to capture the spread of three-dimensional data, and a $N \times N$ covariance matrix captures the spread of N-dimensional data.

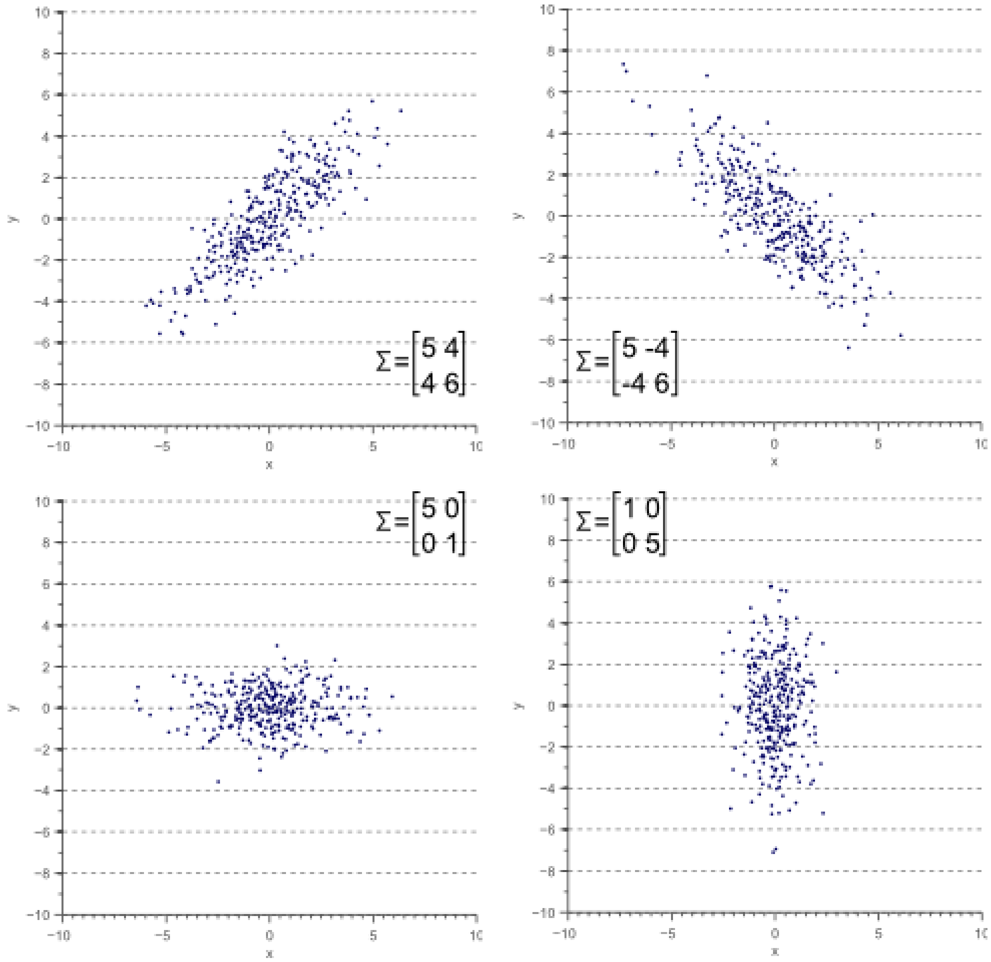


Figure 3. The covariance matrix defines the shape of the data. Diagonal spread is captured by the covariance, while axis-aligned spread is captured by the variance.

Eigendecomposition of a covariance matrix

In the next section, we will discuss how the covariance matrix can be interpreted as a linear operator that transforms white data into the data we observed. However, before diving into the technical details, it is important to gain an intuitive understanding of how eigenvectors and eigenvalues uniquely define the covariance matrix, and therefore the shape of our data.

As we saw in figure 3, the covariance matrix defines both the spread (variance), and the orientation (covariance) of our data. So, if we would like to represent the covariance matrix with a vector and its magnitude, we should simply try to find the vector that points into the direction of the largest spread of the data, and whose magnitude equals the spread (variance) in this direction.

If we define this vector as $\vec{v}$, then the projection of our data D onto this vector is obtained as $\vec{v}^{\intercal} D$, and the variance of the projected data is $\vec{v}^{\intercal} \Sigma \vec{v}$. Since we are looking for the vector $\vec{v}$ that points into the direction of the largest variance, we should choose its components such that the covariance matrix $\vec{v}^{\intercal} \Sigma \vec{v}$ of the is a normalized unit vector, can be formulated as a so-called Rayleigh Quotient. The maximum of such a Rayleigh Quotient is obtained by setting $\vec{v}$ equal to the largest eigenvector of matrix Σ.

In other words, the largest eigenvector of the covariance matrix always points into the direction of the largest variance of the data, and the magnitude of this vector equals the corresponding eigenvalue. The second largest eigenvector is always orthogonal to the largest eigenvector, and points into the direction of the second largest spread of the data.

Now let's have a look at some examples. In an earlier article we saw that a linear transformation matrix T is completely defined by its eigenvectors and eigenvalues. Applied to the covariance matrix, this means that:

$$\Sigma\vec{v} = \lambda\vec{v} \tag{4}$$

where $\vec{v}$ is an eigenvector of Σ, and λ is the corresponding eigenvalue.

If the covariance matrix of our data is a diagonal matrix, such that the covariances are zero, then this means that the variances must be equal to the eigenvalues λ. This is illustrated by figure 4, where the eigenvectors are shown in green and magenta, and where the eigenvalues clearly equal the variance components of the covariance matrix.

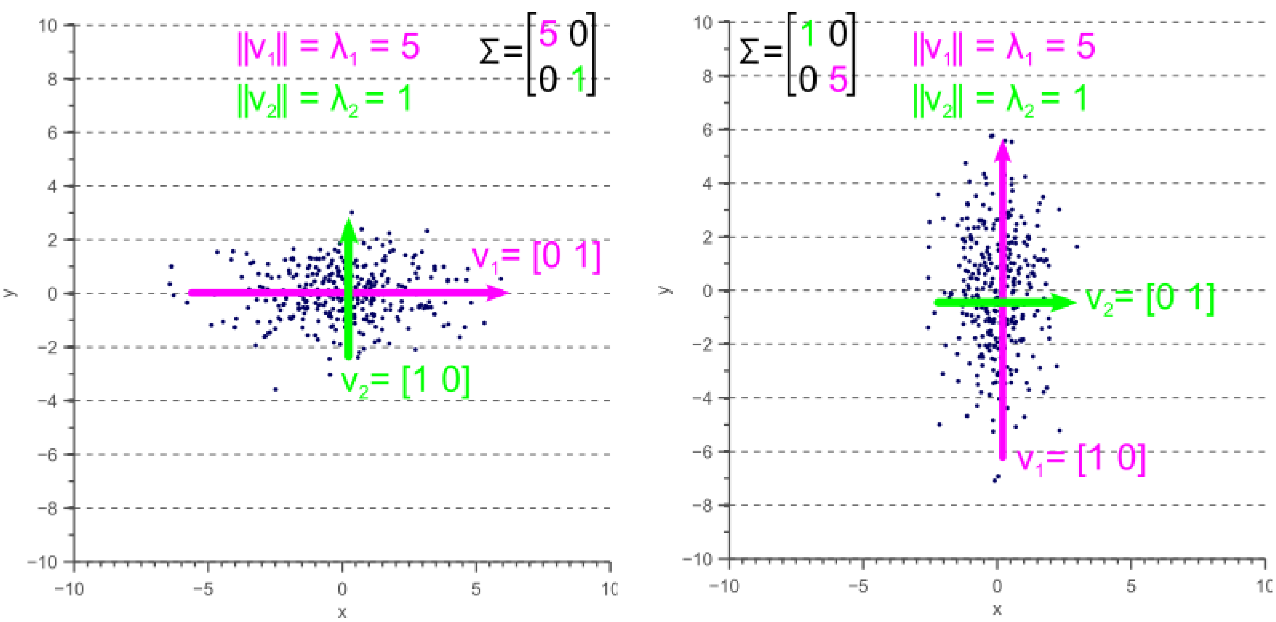


Figure 4. Eigenvectors of a covariance matrix

However, if the covariance matrix is not diagonal, such that the covariances are not zero, then the situation is a little more complicated. The eigenvalues still represent the variance magnitude in the direction of the largest spread of the data, and the variance components of the covariance matrix still represent the variance

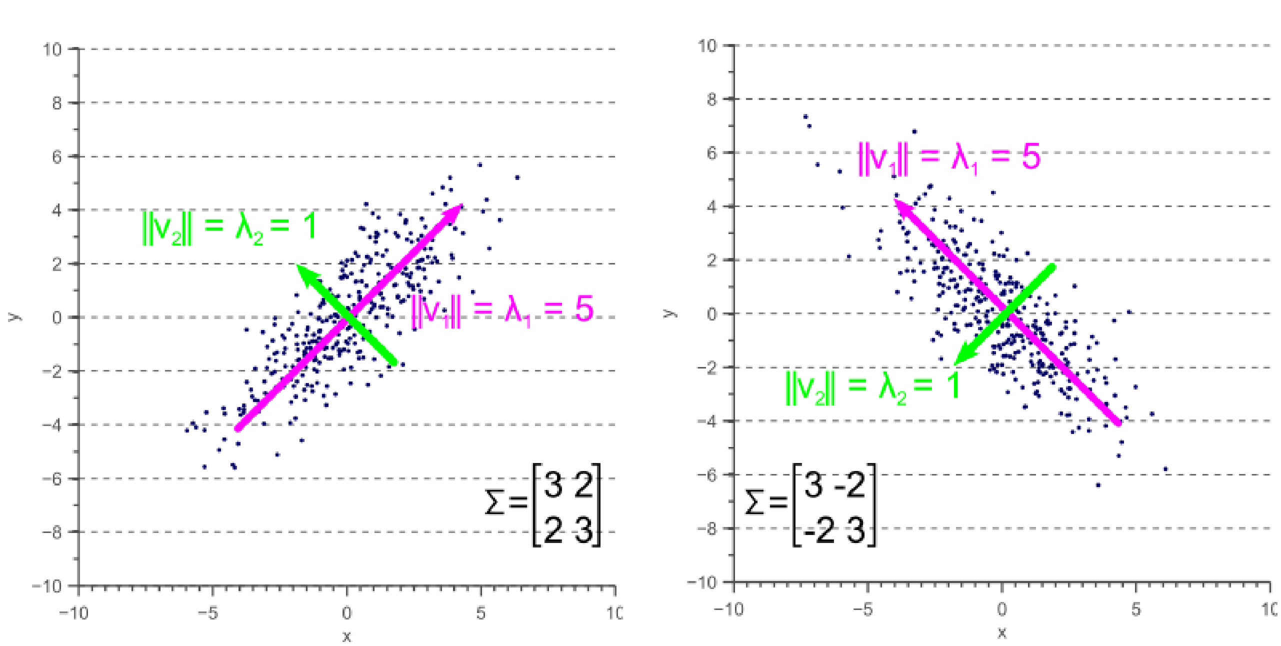


Figure 5. Eigenvalues versus variance

By comparing figure 5 with figure 4, it becomes clear that the eigenvalues represent the variance of the data along the eigenvector directions, whereas the variance components of the covariance matrix represent the spread along the axes. If there are no covariances, then both values are equal.

Covariance matrix as a linear transformation

Now let's forget about covariance matrices for a moment. Each of the examples in figure 3 can simply be considered to be a linearly transformed instance of figure 6:

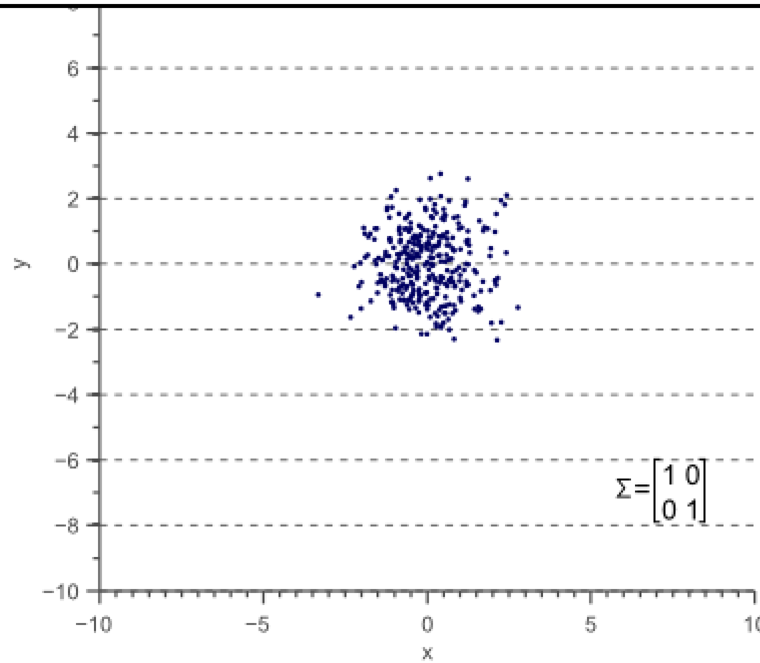


Figure 6. Data with unit covariance matrix is called white data.

Let the data shown by figure 6 be D, then each of the examples shown by figure 3 can be obtained by linearly transforming D:

$$D' = T\,D \tag{5}$$

where T is a transformation matrix consisting of a rotation matrix R and a scaling matrix S:

$$T = R\,S. \tag{6}$$

These matrices are defined as:

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{7}$$

where θ is the rotation angle, and:

$$S = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \tag{8}$$

where s_x and s_y are the scaling factors in the x direction and the y direction respectively.

transformation matrix $T = RS$.

Let's start with unscaled (scale equals 1) and unrotated data. In statistics this is often refered to as 'white data' because its samples are drawn from a standard normal distribution and therefore correspond to white (uncorrelated) noise:

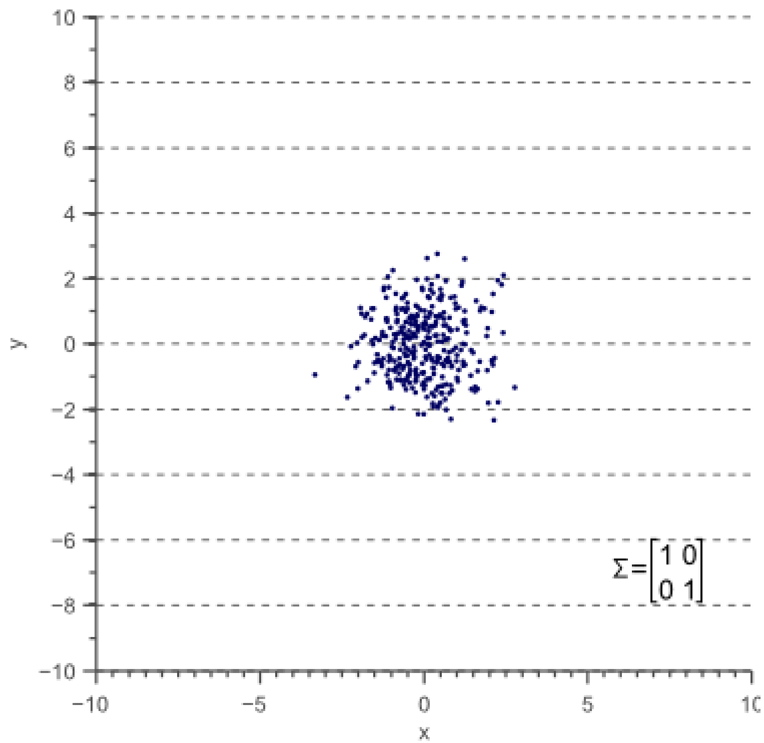


Figure 7. White data is data with a unit covariance matrix.

The covariance matrix of this 'white' data equals the identity matrix, such that the variances and standard deviations equal 1 and the covariance equals zero:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{9}$$

Now let's scale the data in the x-direction with a factor 4:

$$D' = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} D \tag{10}$$

The data D' now looks as follows:

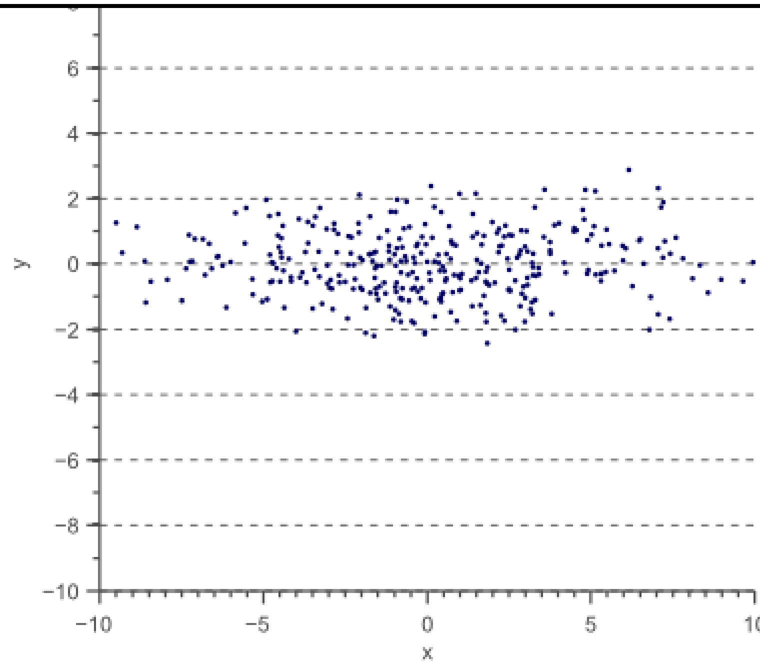


Figure 8. Variance in the x-direction results in a horizontal scaling.

The covariance matrix Σ' of D' is now:

$$\Sigma' = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix} \tag{11}$$

Thus, the covariance matrix Σ' of the resulting data D' is related to the linear transformation T that is applied to the original data as follows: $D' = T\,D$, where

$$T = \sqrt{\Sigma'} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}. \tag{12}$$

However, although equation (12) holds when the data is scaled in the x and y direction, the question rises if it also holds when a rotation is applied. To investigate the relation between the linear transformation matrix T and the covariance matrix Σ' in the general case, we will therefore try to decompose the covariance matrix into the product of rotation and scaling matrices.

As we saw earlier, we can represent the covariance matrix by its eigenvectors and eigenvalues:

$$\Sigma\vec{v} = \lambda\vec{v} \tag{13}$$

Equation (13) holds for each eigenvector-eigenvalue pair of matrix Σ. In the 2D case, we obtain two eigenvectors and two eigenvalues. The system of two equations defined by equation (13) can be represented efficiently using matrix notation:

$$\Sigma V = V L \tag{14}$$

where V is the matrix whose columns are the eigenvectors of Σ and L is the diagonal matrix whose non-zero elements are the corresponding eigenvalues.

This means that we can represent the covariance matrix as a function of its eigenvectors and eigenvalues:

$$\Sigma = V L V^{-1} \tag{15}$$

Equation (15) is called the eigendecomposition of the covariance matrix and can be obtained using a Singular Value Decomposition algorithm. Whereas the eigenvectors represent the directions of the largest variance of the data, the eigenvalues represent the magnitude of this variance in those directions. In other words, V represents a rotation matrix, while $\sqrt{L}$ represents a scaling matrix. The covariance matrix can thus be decomposed further as:

$$\Sigma = R\, S\, S\, R^{-1} \tag{16}$$

where $R = V$ is a rotation matrix and $S = \sqrt{L}$ is a scaling matrix.

In equation (6) we defined a linear transformation $T = R\, S$. Since S is a diagonal scaling matrix, $S = S^{\intercal}$. Furthermore, since R is an orthogonal matrix, $R^{-1} = R^{\intercal}$. Therefore, $T^{\intercal} = (R\, S)^{\intercal} = S^{\intercal}\, R^{\intercal} = S\, R^{-1}$. The covariance matrix can thus be written as:

$$\Sigma = R\, S\, S\, R^{-1} = T\, T^{\intercal}, \tag{17}$$

In other words, if we apply the linear transformation defined by $T = R\, S$ to the original white data D shown by figure 7, we obtain the rotated and scaled data D' with covariance matrix $T\, T^{\intercal} = \Sigma' = R\, S\, S\, R^{-1}$. This is illustrated by figure 10:

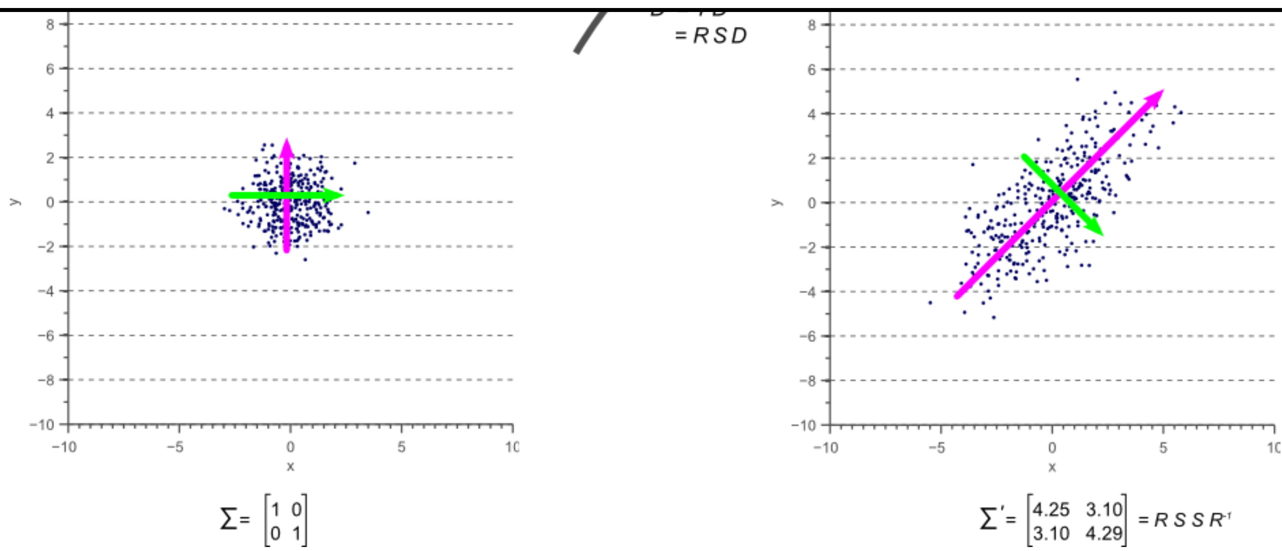


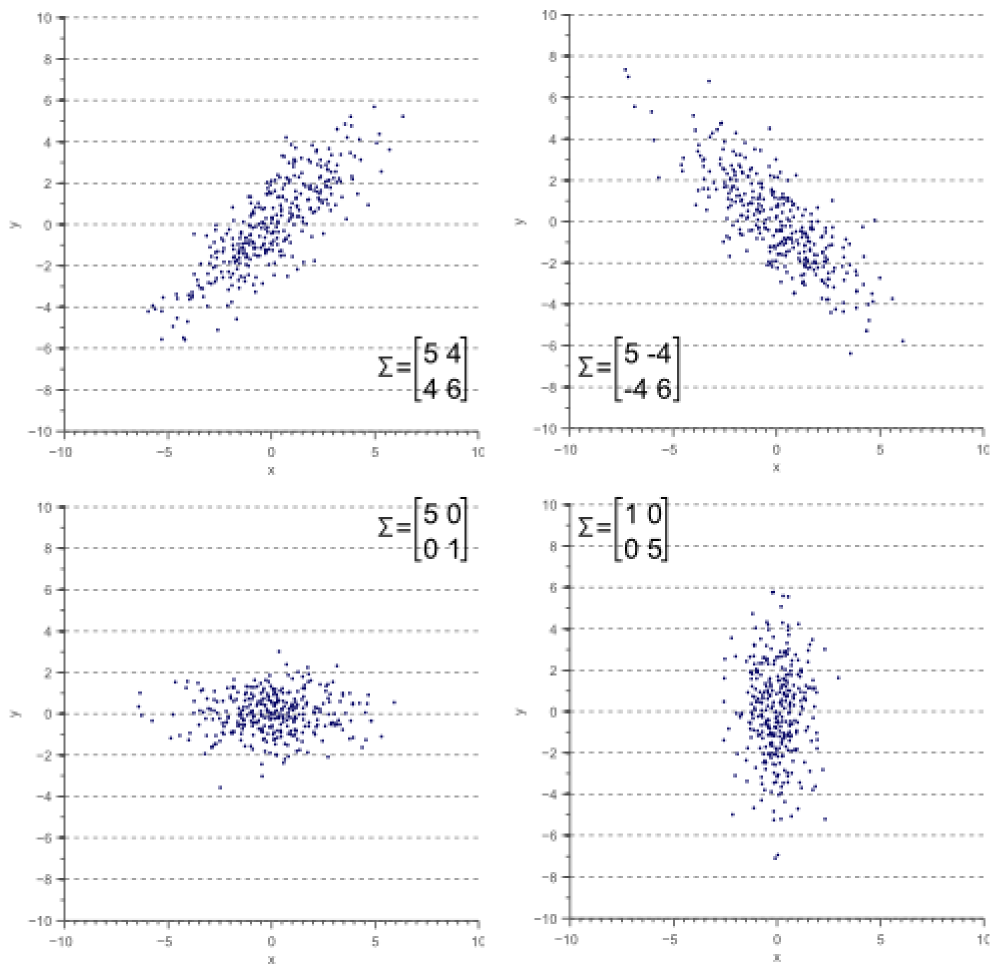
Figure 10. The covariance matrix represents a linear transformation of the original data.

The colored arrows in figure 10 represent the eigenvectors. The largest eigenvector, i.e. the eigenvector with the largest corresponding eigenvalue, always points in the direction of the largest variance of the data and thereby defines its orientation. Subsequent eigenvectors are always orthogonal to the largest eigenvector due to the orthogonality of rotation matrices.

Conclusion

In this article we showed that the covariance matrix of observed data is directly related to a linear transformation of white, uncorrelated data. This linear transformation is completely defined by the eigenvectors and eigenvalues of the data. While the eigenvectors represent the rotation matrix, the eigenvalues correspond to the square of the scaling factor in each dimension.

If you're new to this blog, don't forget to subscribe, or follow me on twitter!



Article Name	A geometric interpretation of the covariance matrix
Author	Vincent Spruyt
Description	In this article, we provide an intuitive, geometric interpretation of the covariance matrix, by exploring the relation between linear transformations and the resulting data covariance.

April 24, 2014 Vincent Spruyt Linear algebra 47 Comments covariance matrix, eigendecomposition, Eigenvectors, linear transformation, PCA

Comments

Chris says:

May 14, 2014 at 3:42 pm

Great article thank you

Reply

Alex says:

May 14, 2014 at 8:09 pm

The covariance matrix is symmetric. Hence we can find a basis of orthonormal eigenvectors and then $\Sigma=VL V^T$.

From computational point of view it is much simpler to find V^T than V^{-1}.

Reply

Vincent Spruyt says:

May 15, 2014 at 7:59 am

This is also written in the article: “Furthermore, since R is an orthogonal matrix, R^{-1} = R^T”. But you are right that I only mention this near the end of the article, mostly because it is easier to develop an intuitive understanding of the first part of the article by considering R^{-1} instead of R^T.

Reply

Brian says:

May 24, 2014 at 1:03 am

Great post! I had a couple questions:

1) The data D doesn’t need to be Gaussian does it?

2) Is [9] reversed (should D be on the left)?

Reply

Vincent Spruyt says:

May 24, 2014 at 8:27 am

Hi Brian:

1) Indeed the data D does not need to be

have made that more clear in the article. However, talking about covariance matrices often does not have much meaning in highly non-Gaussian data.

2) That depends on whether D is a row vector or a column vector I suppose. In this case, if each column of D is a data entry, then R*D = (D^t*R)^t

Reply

Konstantin says:

August 18, 2014 at 1:47 am

Thank you for this great post! But let me please correct one fundamental mistake that you made. The square root of covariance matrix M is not equal to R * S. The square root of M equals R * S * R', where R' is transposed R. Proof: (R * S * R') * (R * S * R') = R * S * R' * R * S * R' = R * S * S * R' = T * T' = M. And, of course, T is not a symmetric matrix (in your post T = T', which is wrong).

Reply

August 18, 2014 at 8:29 am

Thanks a lot for noticing! You are right indeed, I will get back about this soon (don't really have time right now).

Edit: I just fixed this mistake. Sorry for the long delay, I didn't find the time before. Thanks a lot for your feedback!

Reply

srinivas kumar says:

October 29, 2014 at 2:52 pm

Very Useful Article 😃 What I feel needs to be included is the interpretation of the action of the covariance matrix as a linear operator. For example, the eigen vectors of the covariance matrix form the principal components in PCA. So, basically , the covariance matrix takes an input data point (vector) and if it resembles the data points from which the operator was obtained, it keeps it invariant (

eigenvectors of covariance matrix ?

Reply

Vincent Spruyt says:
March 7, 2015 at 2:18 pm

Hi Kumar, great point! This is basically captured by equations 13 and 14, but I just added a short section to make this a bit more clear in the article.

Reply

Srinivas Kumar says:
March 7, 2015 at 3:02 pm

"..the eigenvectors represent the directions of the largest variance of the data, the eigenvalues represent the magnitude of this variance in those directions.." … 😃 Thanks a lot for expressing it so precisely.

Reply

Paul says:

April 25, 2015 at 5:40 pm

Thank you for such an intuitive article. I have spent countless hours over countless days trying to picture exactly what you described. I wonder if you can clarify something in the writing, though. When you first talk about vector v, throughout the entire paragraph, it is referred to both as a unit vector and a vector whose length is set to match the spread of data in the direction of v.

Reply

Paul says:

April 25, 2015 at 5:53 pm

By the way, would you know of a similarly intuitive description of cov(X,Y), where X and Y are disjoint sets of random variables?

Reply

Manoj Nambiar says:

May 14, 2015 at 5:30 am

its components such that the covariance matrix \vec{v}^{\intercal} \Sigma \vec{v} of the projected data is as large as possible….”

That quantity “\vec{v}^{\intercal} \Sigma \vec{v}” (sorry - I am not able to do a graphical paste - but I hope you know what I mean) is not a matrix - It is a scalar quantity - isn’t it?

Or what you wanted to say was ” …..we should choose its components such that the covariance of the data with the vector v is as large as possible….”. And this covariance is a term of the Raliegh’s coefficient ……

May be there is a better way to put …

Reply

harmyder says:

January 17, 2017 at 9:30 am

The better way is to say that it is just a variance of projected data. So, that is a mistake, it should be variance, not covariance.

Lukas says:

June 2, 2015 at 5:09 pm

Hello Vincent,

Thank you very much for this blog post. I have one question though concerning figure 4: Shouldn't the magenta eigenvector in the right part of the picture point downwards? Otherwise, we wouldn't have a proper rotation.

Best regards,
Lukas

Reply

Gordon Marney says:

June 8, 2015 at 12:59 am

The correlations you showin figure 5 look a lot like Reduced Major Axis. How do do the eigenvectors and RMA compare?. Of course, there is nothing like eigenvalues in values after rotation of the RMA regression?

Gordon

Reply

Inderjit Nanda says:
June 9, 2015 at 12:49 pm

I love to reread your articles. Hope to see more such intuitive topics!!!

Reply

Mehdi Pedram says:
June 12, 2015 at 4:24 pm

Great Thank you.

Reply

seravee says:
June 17, 2015 at 9:09 am

Thanks a lot. Very intuitive articles on the covariance matrix.

Pradeep says:

June 21, 2015 at 10:31 am

Always wondered why Eigen vectors of covariance matrix and the actual data were similar. Thanks for the tutorial. It helped in clearing the doubt.

Reply

haining says:

September 11, 2015 at 6:36 pm

Great article!!! It's soooooo helpful, thank you 🙂

Reply

Priyamvad says:

September 13, 2015 at 11:52 pm

Thanks for sharing this article, it's a wonderful read!
Am I correct in understanding that the transformation TT^t for the covariance matrix will apply when transforming any data by T, not just for white data?

Reply

Joe says:

October 30, 2015 at 2:21 pm

Nice article. Thank you so much!

Reply

Jim Price says:

December 1, 2015 at 4:03 pm

Superb!

It is just awesome that you are so open to suggestions and then make the changes for the benefit of all of us.

Thank You, Thank You, Thank You!!

Reply

dk sunil says:

January 18, 2016 at 9:59 pm

excellent article. gave me a whole new perspective of covariance matrix.

Reply

Gennaro says:

Thanks man, great article! So useful for my PhD !

Do you know of any mathematics book where I can find a rigorous dissertation about this? My professor wants me to be as meticulous as possible 🙁

Thank you in advance!

Reply

simon says:

March 10, 2017 at 6:11 am

Do you know of any mathematics book where I can find a rigorous dissertation about this?

Reply

Nariman says:

February 12, 2016 at 2:51 pm

education system need people like you

Reply

beena says:

March 20, 2016 at 6:22 pm

Reply

Marcell says:

March 29, 2016 at 6:56 pm

This is a great article, thank you so much! I completely agree with your motivation to write things down in a simple way, rather than trying to sound smart to people who already know everything. I learned so much from your blog in a short time.

Reply

Treble says:

April 15, 2016 at 2:16 am

Thanks for this! This and the companion eigen decomposition article were exactly what I was looking for, and much easier to understand than other resources I found.

Reply

Artur RC says:

Great text. Good job! Thank you 😐

Reply

Shibumon Alampatta says:

April 21, 2016 at 5:56 am

Great article, but one doubt. It was mentioned that direction of eigen vector remains unchanged when linear transformation is applied. But in Fig. 10, direction of the vector is also changed. Can you please explain it?

Reply

ninjajack says:

April 28, 2016 at 7:20 am

Very useful. Thanks!

Reply

Manoj Gupta says:

March 6, 2017 at 10:41 am

Very good explain and worthful.

But I have doubt why does eigenvector have one

It’s true that both cancel out and we are left with zero…
Where I am going geometrically worng.
Second
Why don’t we a complex eigen vector conjugate when we rotate the white data by rotational matrix….

Reply

Thomas says:
July 13, 2016 at 10:39 am

Really intuitive write up, it was a joy to read.

In figures 4 and 5, though, the v_i are unit vectors and have norm 1. You want to write var(v_i) instead of norm(v_i) in both those figures, as norm(v_i) = 1 but var(v_i) = λ_i.

Reply

Summer says:
July 28, 2016 at 7:51 pm