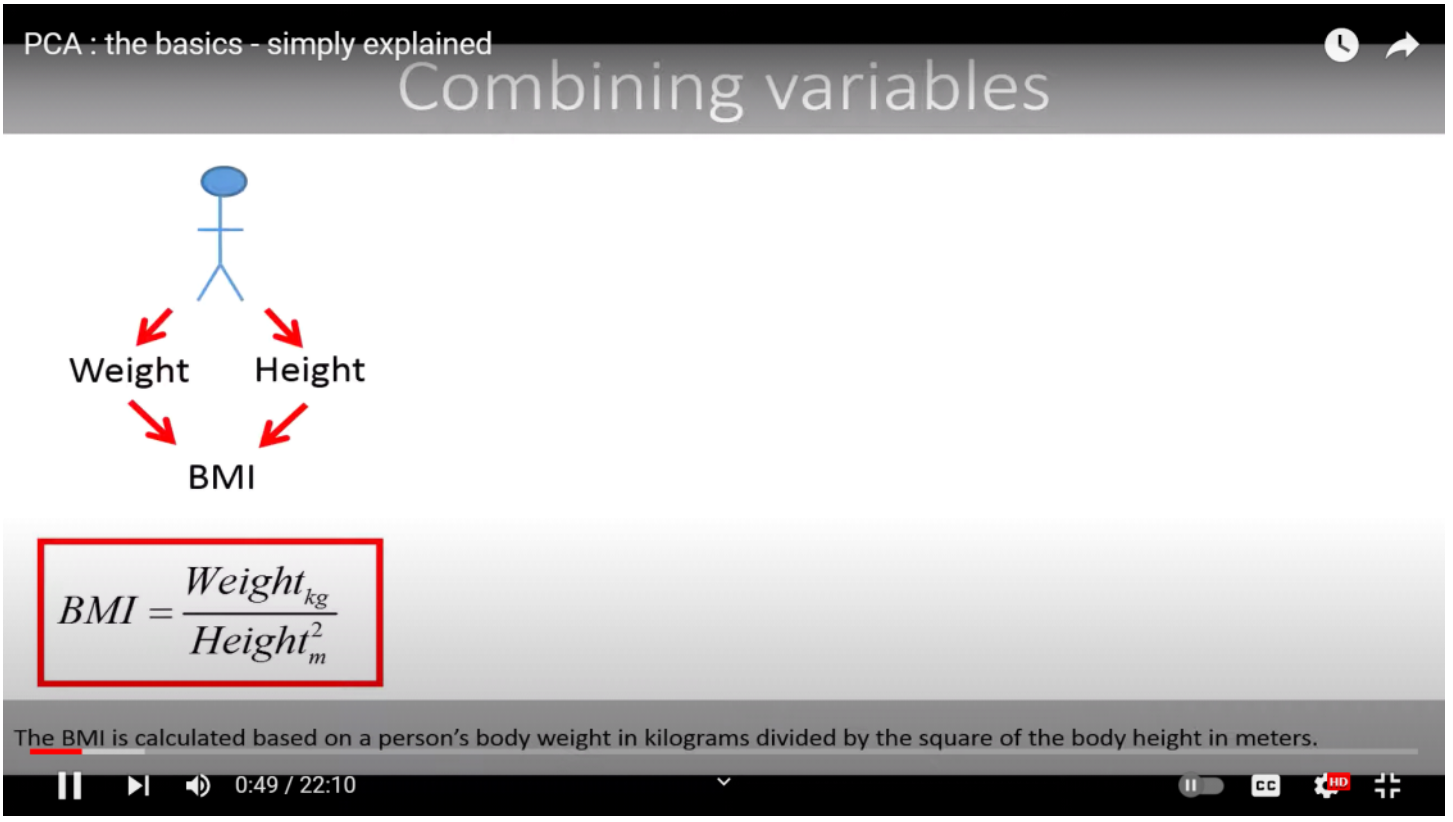
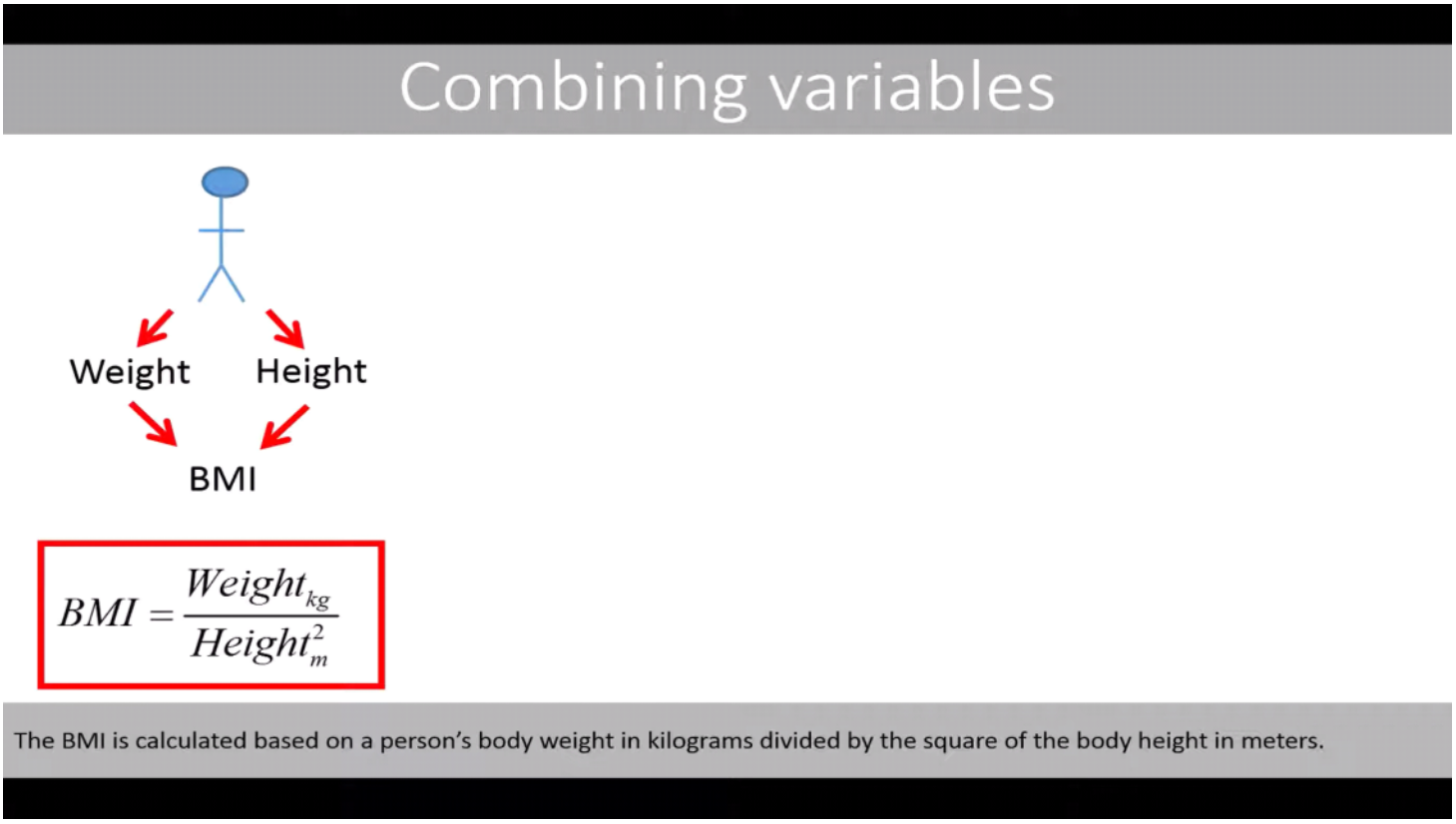
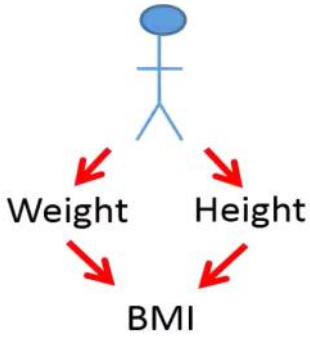
PCA BASICS

10 October 2023 10:39





Combining variables

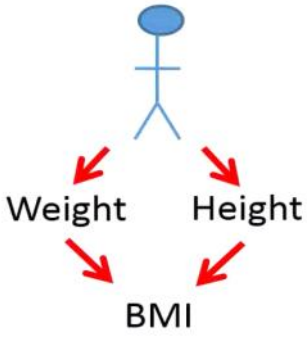


Cholesterol = Weight + Height

$$BMI = \frac{Weight_{kg}}{Height_{m}^{2}}$$

For example, let's say that we like to predict the cholesterol level based on a person's weight and height by using linear regression.

Combining variables

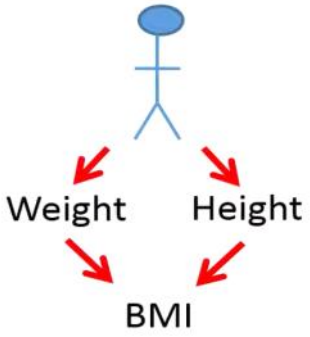


Cholesterol = Weight + Height

$$BMI = \frac{Weight_{kg}}{Height_{m}^{2}}$$

However, remember that it could be problematic in linear regression if there is a too strong correlation between the explanatory variables, which is called multicollinearity.

Combining variables

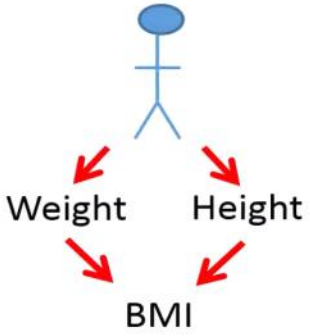


Cholesterol = Weight + Height

$$BMI = \frac{Weight_{kg}}{Height_{m}^{2}}$$

Also, the more explanatory variables we have in our model, the more measurements we need to do. Thus, the sample size generally needs to be bigger for models including more explanatory variables.

Combining variables



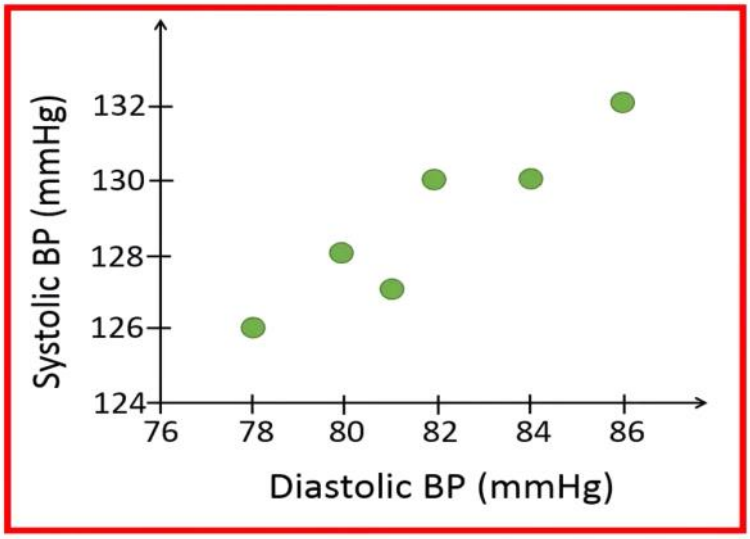
Cholesterol = Weight + Height

Cholesterol = BMI

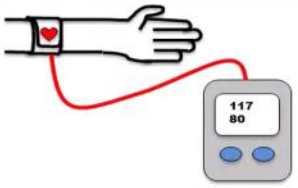
$$BMI = \frac{Weight_{kg}}{Height_{m}^{2}}$$

This is one of the reasons why it makes sense to combine weight and height into just one variable, the body mass index. We can then predict the cholesterol level with just one variable that contains information on both weight and height.

Combining variables

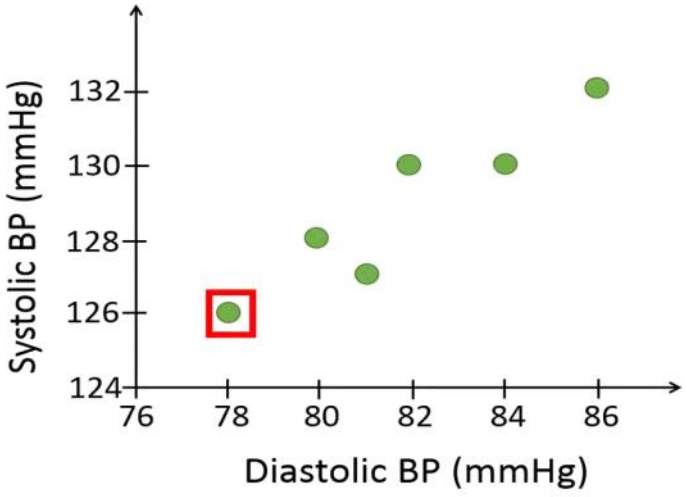


Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

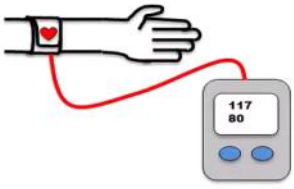


Let's have a look at another example. In this case, we have measured the upper and lower blood pressure of six individuals.

Combining variables

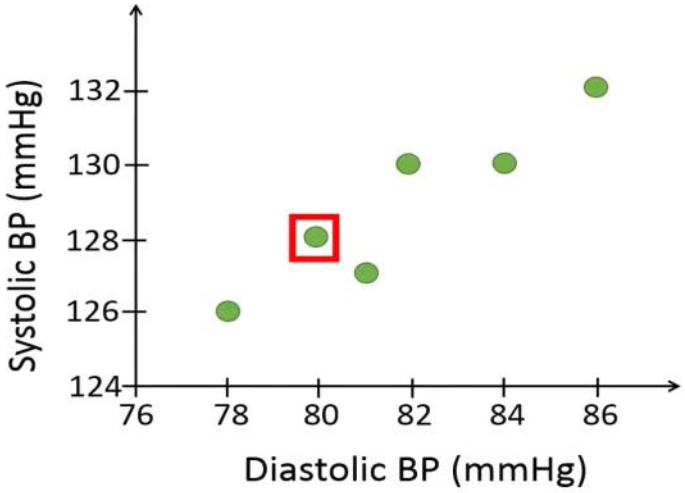


Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

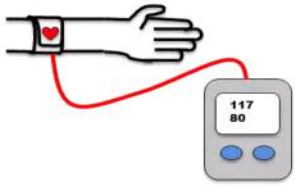


For example, person number one has a diastolic blood pressure of 78 and a systolic blood pressure of 126.

Combining variables

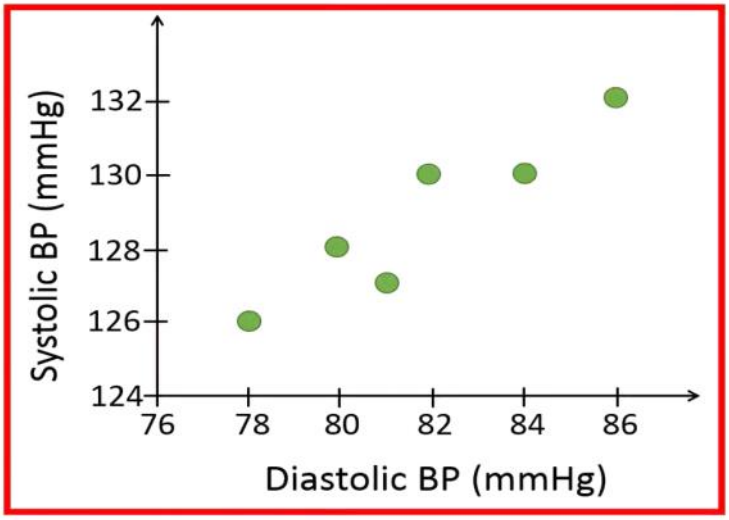


Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

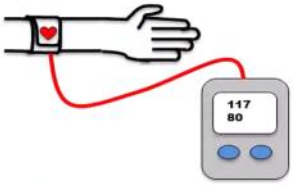


Person number two has a diastolic blood pressure of 80 and a systolic blood pressure of 128, and so on.

Combining variables

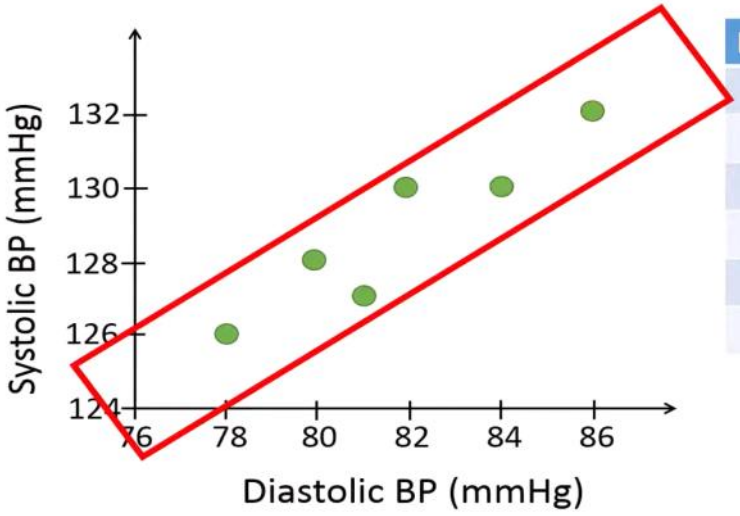


Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

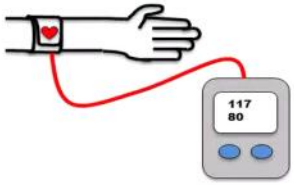


For this data set, there seems to be a strong positive correlation between the upper and lower blood pressure. If a person has a high systolic blood pressure, it is likely that the person also has a high diastolic blood pressure.

Combining variables



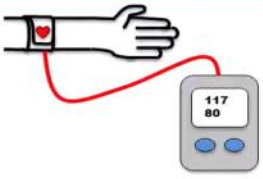
Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



Note that, PCA will be more useful when the variables are strongly correlated, because the combined variable will then contain more information of the variables compared to if the variables show a weak correlation to each other.

Combining variables

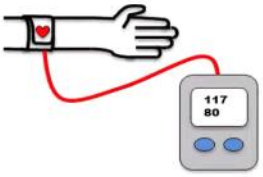
Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



Let's say that we like to combine the upper and lower blood pressure into just one variable that we simply call just blood pressure (BP). However, how do we combine these two variables in the best way?

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

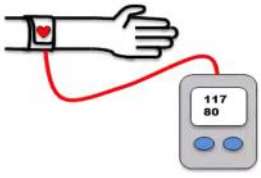


$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

We could use the following equation to combine the two variables,

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

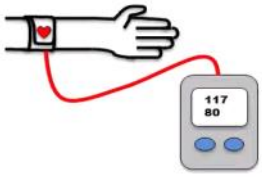


$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

where X1 and X2 represent the two variables that we like to combine.

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



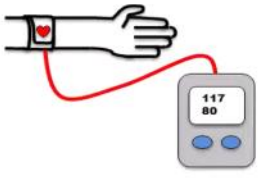
$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

Let's rename X1 and X2 to our measured variables, the diastolic blood pressure (DBP) and the systolic blood pressure (SBP).

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



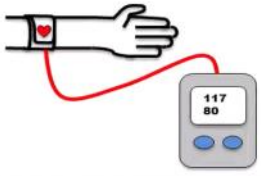
$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

We call the combined variable blood pressure.

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



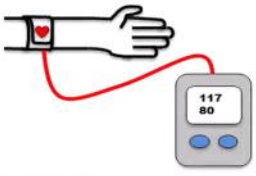
$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

Alpha one and alpha two are called weights. In PCA, these weights are usually referred to as loadings.

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

$$BP = 0.8\,DBP + 0.6\,SBP$$

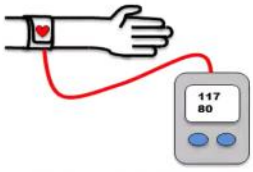
For example, if we set alpha one to 0.8 and alpha two to 0.6,

Combining variables

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

$$BP = 0.8\,DBP + 0.6\,SBP$$

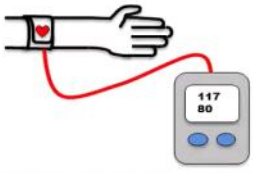


then we put more weight on the diastolic blood pressure than on the systolic blood pressure. This means that the combined variable will be based on more information from the diastolic blood pressure.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$



Let's say that we like to combine the two variables by calculating the mean of the measurements for each individual.

PCA : the basics - simply explained

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = \frac{DBP_1 + SBP_1}{2}$$

For example, to calculate the combined blood pressure for person number one, we add the diastolic and systolic blood pressure, and divide by two since we have two variables in this case.

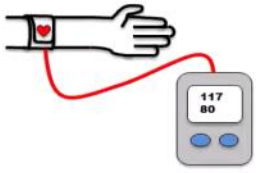
4:20 / 22:10

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = \frac{DBP_1 + SBP_1}{2} = \frac{78 + 126}{2}$$



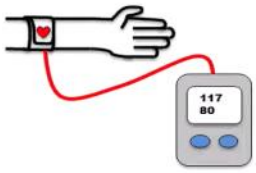
78 plus 126 divided by two is,

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = \frac{DBP_1 + SBP_1}{2} = \frac{78 + 126}{2} = 102$$



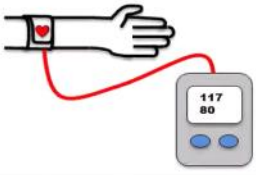
102.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

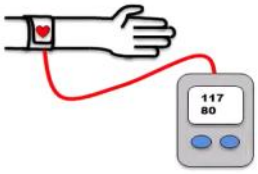
$$BP_1 = \frac{DBP_1 + SBP_1}{2}$$



Note that we can reformulate this equation to,

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	



$$BP = \alpha_1 DBP + \alpha_2 SBP$$

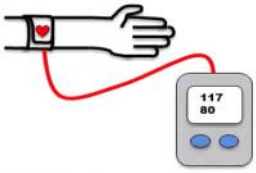
$$BP_1 = \frac{DBP_1 + SBP_1}{2}$$

$$BP_1 = \frac{1}{2} DBP_1 + \frac{1}{2} SBP_1$$

this where we instead multiply by one half before we add the numbers.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	



$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = \frac{DBP_1 + SBP_1}{2}$$

$$BP_1 = \frac{1}{2} DBP_1 + \frac{1}{2} SBP_1$$

Instead of one over two,

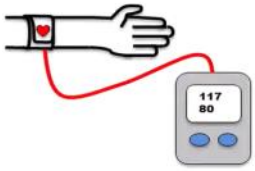
Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = \frac{DBP_1 + SBP_1}{2}$$

$$BP_1 = 0.5 DBP_1 + 0.5 SBP_1$$



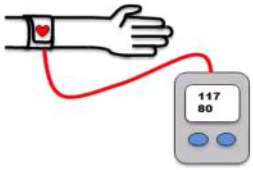
we can multiply by 0.5.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = 0.5 DBP_1 + 0.5 SBP_1$$



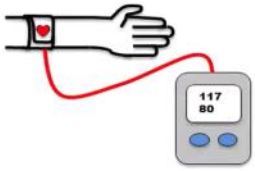
Note that the form of this equation is

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = 0.5DBP_1 + 0.5SBP_1$$



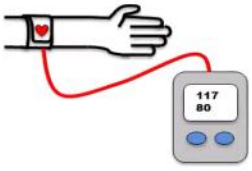
identical to this one.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = 0.5DBP_1 + 0.5SBP_1$$



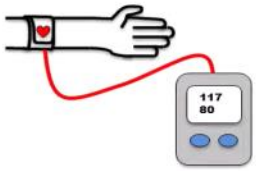
Thus, when we combine the two variables by using the mean, this can be seen as we use the weights 0.5 for our linear combination. When we use this method, we put equal weights on the two variables when we combine them.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_1 = 0.5 DBP_1 + 0.5 SBP_1$$



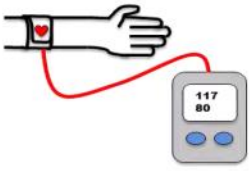
Thus, when we combine the two variables by using the mean, this can be seen as we use the weights 0.5 for our linear combination. When we use this method, we put equal weights on the two variables when we combine them.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	104
81	127	
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_2 = 0.5 \cdot 80 + 0.5 \cdot 128 = 104$$



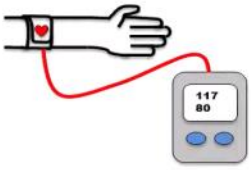
We then continue to combine the values for person number two, by using the same equation with the same weights.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	104
81	127	104
82	130	
84	130	
86	132	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP_3 = 0.5 \cdot 81 + 0.5 \cdot 127 = 104$$

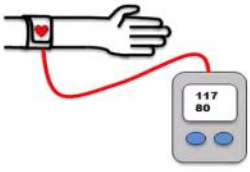


Then for person number three and so on.

Combining variables

Diastolic BP	Systolic BP	Mean BP
78	126	102
80	128	104
81	127	104
82	130	106
84	130	107
86	132	109

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

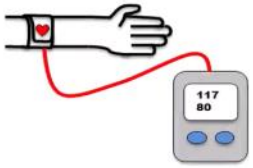


By using this method, we have combined the two variables into just one, by using the mean of the two measurements.

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	
80	128	104	
81	127	104	
82	130	106	
84	130	107	
86	132	109	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

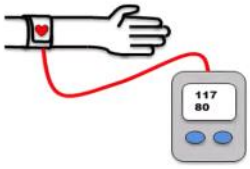


Another way to combine the variables is to simply sum the two measurements for each person.

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	204
80	128	104	
81	127	104	
82	130	106	
84	130	107	
86	132	109	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

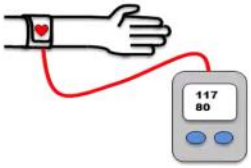


For example, 78 plus 126 is 204,

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	204
80	128	104	208
81	127	104	
82	130	106	
84	130	107	
86	132	109	

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

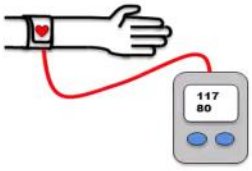


80 plus 128 is 208, and so forth.

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	204
80	128	104	208
81	127	104	208
82	130	106	212
84	130	107	214
86	132	109	218

$$BP = \alpha_1 DBP + \alpha_2 SBP$$



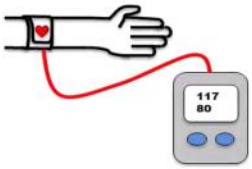
By using the sum, these values represent our combined variable.

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	204
80	128	104	208
81	127	104	208
82	130	106	212
84	130	107	214
86	132	109	218

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP = 1 \cdot DBP + 1 \cdot SBP$$



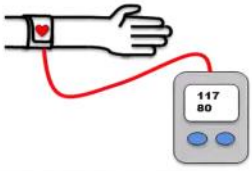
Note that, when we sum the values, we also use the same basic formula as we used when we combined the variables based on the mean.

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	204
80	128	104	208
81	127	104	208
82	130	106	212
84	130	107	214
86	132	109	218

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

$$BP = 1 \cdot DBP + 1 \cdot SBP$$

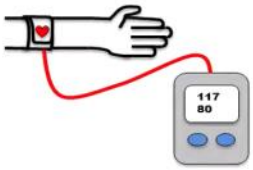


The difference is that we here set the weights to one instead of 0.5.

Combining variables

Diastolic BP	Systolic BP	Mean BP	Sum BP
78	126	102	204
80	128	104	208
81	127	104	208
82	130	106	212
84	130	107	214
86	132	109	218

$$BP = \alpha_1 DBP + \alpha_2 SBP$$



In conclusion, the two methods that we have used so far use the same equation to combined the two variables. The difference between the two methods is just the values used for the weights. We will now discuss principal component analysis.

PCA

Principal component analysis (PCA) is a method to find the linear combination that accounts for as much variability as possible.

PCA

Principal component analysis (PCA) is a method to find the linear combination that accounts for as much variability as possible.

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

In our example, PCA would combine the diastolic blood pressure and the systolic blood pressure in a way,

PCA

Principal component analysis (PCA) is a method to find the linear combination that accounts for as much variability as possible.

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

so that the combined variable has as much variability as possible. In other words, it will combine the two variables so that we maximize the variance of the combined variable.

PCA

Principal component analysis (PCA) is a method to find the linear combination that accounts for as much variability as possible.

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

PCA therefore tries to find the optimal values for alpha one and alpha two that maximize the variance of the linear combination.

PCA

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

Since one can make the variance larger by simply selecting larger values for the weights,

PCA

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

Constraint:

$$\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_p^2 = 1$$

the basic PCA therefore uses the following constraint, where the squared alpha values should sum up to one. PCA also uses other types of constraints that we will discuss in another video.

PCA

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

Constraint:

$$\alpha_1^2 + \alpha_2^2 = 1$$

In our example, we want to combine only two variables, which means that the square of alpha one plus the square of alpha two should be equal to one.

PCA

$$BP = \alpha_1 DBP + \alpha_2 SBP$$

Constraint:

$$\alpha_1^2 + \alpha_2^2 = 1$$

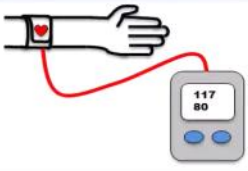
$$0.8^2 + 0.6^2 = 0.64 + 0.36 = 1$$

For example, if we set alpha one to 0.8 and alpha two to 0.6, we see that the sum of the squared values is equal to one.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	
80	128	
81	127	
82	130	
84	130	
86	132	



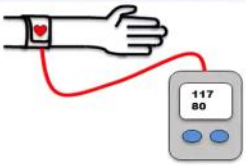
Let's use these weights in order to combine the two variables.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	138.0
80	128	
81	127	
82	130	
84	130	
86	132	

$$BP = 0.8 \cdot 78 + 0.6 \cdot 126 = 138$$



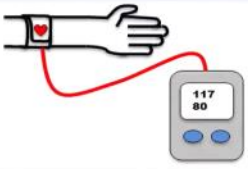
By using the linear combination of the two variables with the weights 0.8 and 0.6, the first person has a combined blood pressure of 138.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	138.0
80	128	140.8
81	127	
82	130	
84	130	
86	132	

$$BP = 0.8 \cdot 80 + 0.6 \cdot 128 = 140.8$$



The second person has a value of 140.8, and so forth.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	138.0
80	128	140.8
81	127	141.0
82	130	143.6
84	130	145.2
86	132	148.0

Next, we calculate the variance of this combined variable.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	138.0
80	128	140.8
81	127	141.0
82	130	143.6
84	130	145.2
86	132	148.0
	Mean =	**142.8**

We therefore first need to calculate the mean.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	138.0
80	128	140.8
81	127	141.0
82	130	143.6
84	130	145.2
86	132	148.0
	Mean =	**142.8**

$$\text{var}(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{5}\left((138-142.8)^2 + \ldots + (148-142.8)^2\right) = 12.74$$

Remember that the sample variance is calculated as the sum of the squared difference between the individual values and the mean, divided by the sample size minus one.

PCA

$$BP = 0.8DBP + 0.6SBP$$

Diastolic BP	Systolic BP	BP
78	126	138.0
80	128	140.8
81	127	141.0
82	130	143.6
84	130	145.2
86	132	148.0
	Mean =	**142.8**

$$\text{var}(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{5}\left((138-142.8)^2 + \ldots + (148-142.8)^2\right) = 12.74$$

In this case, we see that variance of the combined variable is 12.74.

PCA

α_1	α_2	var(Y)
0.8	0.6	12.74

$$BP = 0.8DBP + 0.6SBP$$

Let's try many different weights for our linear combination to see which combination that results in the maximum variance of the combined variable.

PCA

α_1	α_2	var(Y)
0.8	0.6	12.74

$$BP = 0.8DBP + 0.6SBP$$

From our previous example, we know that if the weights are set to 0.8 and 0.6, the combined variable has a variance of 12.74.

PCA

α_1	α_2	var(Y)	
0.8	0.6	12.74	$BP = 0.8DBP + 0.6SBP$
0.6	0.8	11.8	$BP = 0.6DBP + 0.8SBP$

If we flip the values of the weights, so that we put more weight on the systolic blood pressure, the variance is reduced to 11.8. This indicates that we should put more weight on the diastolic blood pressure to maximize the variance.

PCA

α_1	α_2	var(Y)	
0.8	0.6	12.74	$BP = 0.8DBP + 0.6SBP$
0.6	0.8	11.8	$BP = 0.6DBP + 0.8SBP$
0.98	0.2	10.4	$BP = 0.98DBP + 0.2SBP$

Let's put a lot of weight on the diastolic blood pressure with the weights 0.98 and 0.2. The sum of these squared weights is approximately equal to one.

PCA

α_1	α_2	var(Y)
0.8	0.6	12.74
0.6	0.8	11.8
0.98	0.2	10.4

$$BP = 0.8DBP + 0.6SBP$$
$$BP = 0.6DBP + 0.8SBP$$
$$BP = 0.98DBP + 0.2SBP$$

By putting too much weight on the diastolic blood pressure, we reduce the variance of the combined variable to 10.4.

PCA

α_1	α_2	var(Y)
0.8	0.6	12.74
0.6	0.8	11.8
0.98	0.2	10.4
0.2	0.98	7.4

$$BP = 0.8DBP + 0.6SBP$$
$$BP = 0.6DBP + 0.8SBP$$
$$BP = 0.98DBP + 0.2SBP$$
$$BP = 0.2DBP + 0.98SBP$$

Finally, we test to put a lot of weight on the systolic blood pressure instead. These weights result in the lowest variance of the combined variable out of all weights we have tried so far.

PCA

α_1	α_2	var(Y)
0.8	0.6	12.74
0.6	0.8	11.8
0.98	0.2	10.4
0.2	0.98	7.4

$$BP = 0.8DBP + 0.6SBP$$
$$BP = 0.6DBP + 0.8SBP$$
$$BP = 0.98DBP + 0.2SBP$$
$$BP = 0.2DBP + 0.98SBP$$

According to our basic analysis, we would select the weights 0.8 and 0.6 when we combine the diastolic and systolic blood pressures because these weights generate maximal variance of the combined variable.

PCA

α_1	α_2	var(Y)
0.8	0.6	12.74
0.6	0.8	11.8
0.98	0.2	10.4
0.2	0.98	7.4

These are the fundamental basics behind PCA. It finds the optimal values of the weights in order to maximize the variance of the combined variable. PCA puts different weights on the variables that are combined to maximize the variance.

PCA

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

$$BP = 0.8DBP + 0.6SBP$$

So how does the PCA find the optimal weights? In the next lecture, we will look into the mathematical details. For now, we keep things simple and explain the method very briefly.

PCA

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

	DBP	SBP
DBP	8.17	5.97
SBP	5.97	4.97

$$BP = 0.8DBP + 0.6SBP$$

In our example, the PCA would first compute the following covariance matrix.

PCA

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

	DBP	SBP
DBP	8.17	5.97
SBP	5.97	4.97

$$Eig = \begin{bmatrix} -0.8 \\ -0.6 \end{bmatrix}$$

$$BP = -0.8DBP + (-0.6SBP)$$

Then it would compute the two eigenvectors of the covariance matrix. The eigenvector with the largest eigenvalue holds the values of the weights, which in this case are minus 0.8 and minus 0.6.

PCA

Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

	DBP	SBP
DBP	8.17	5.97
SBP	5.97	4.97

$$Eig = \begin{bmatrix} -0.8 \\ -0.6 \end{bmatrix}$$

$$BP = -0.8DBP + (-0.6SBP)$$

The values in the first eigenvector are then used as weights to combine the two variables. Although the weights are negative in this case, we will get the same variance as if they would have been positive as in our previous example.

PCA

Applications

We will now have a look at some examples where PCA is used to analyze biological data.

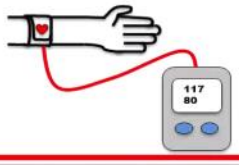
PCA

Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164

PCA can be used to reduce the number of dimensions or variables in our data set for further types of analysis. We will here see how we can reduce the following four variables into just two variables.

PCA

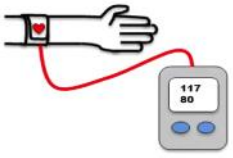
Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164



In this example data set, one has measured the diastolic and systolic blood pressure of six individuals,

PCA

Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164



as well as their body weight in kilos and body height in centimeters.

PCA

Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164

BP BS

For example, we could use PCA to combine these four variables into two new variables.

PCA

Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164

BP BS

If the diastolic and systolic blood pressure show a strong correlation, we could combine them into just one variable that we call blood pressure,

PCA

Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164

BP

BS

and if weight and height show a strong correlation, we could combine them into a variable that we call body size. In later videos we will see how we can perform this kind of combination.

PCA

Diastolic BP	Systolic BP	Weight	Height
78	126	67	170
80	128	77	177
81	127	89	183
82	130	90	187
84	130	50	165
86	132	55	164

BP

BS

Cholesterol = BP + BS

If we would use these variables to predict, for example, the cholesterol level with linear regression, we could use only two explanatory variables, the blood pressure and the body size.

PCA

Person	DBP	SBP	BMI	Chol.	Pulse	Temp	...
1	78	126	25	170	55	37.4	...
2	80	128	27	177	56	37.8	...
3	81	127	23	183	60	36.8	...
4	82	130	30	187	61	36.4	...
5	84	130	28	165	62	36.9	...
6	86	132	35	164	70	37.0	...
...	..	..	..	..	..	..	...

We will now have a look at another example. Let's say that we have measured many clinical variables on many individuals.

PCA

Person	DBP	SBP	BMI	Chol.	Pulse	Temp	...
1	78	126	25	170	55	37.4	...
2	80	128	27	177	56	37.8	...
3	81	127	23	183	60	36.8	...
4	82	130	30	187	61	36.4	...
5	84	130	28	165	62	36.9	...
6	86	132	35	164	70	37.0	...
...	..	..	..	..	..	..	...

Suppose we like to identify people that have a similar health profile, which means that they have about the same values of the clinical variables that have been measured.

PCA

Person	DBP	SBP	BMI	Chol.	Pulse	Temp	...
1	78	126	25	170	55	37.4	...
2	80	128	27	177	56	37.8	...
3	81	127	23	183	60	36.8	...
4	82	130	30	187	61	36.4	...
5	84	130	28	165	62	36.9	...
6	86	132	35	164	70	37.0	...
...	..	..	..	..	..	..	...

For example, is the health profile of person number 1 more similar to

PCA

Person	DBP	SBP	BMI	Chol.	Pulse	Temp	...
1	78	126	25	170	55	37.4	...
2	80	128	27	177	56	37.8	...
3	81	127	23	183	60	36.8	...
4	82	130	30	187	61	36.4	...
5	84	130	28	165	62	36.9	...
6	86	132	35	164	70	37.0	...
...	..	..	..	..	..	..	...

the health profile of person number two,

PCA

Person	DBP	SBP	BMI	Chol.	Pulse	Temp	...
1	78	126	25	170	55	37.4	...
2	80	128	27	177	56	37.8	...
3	81	127	23	183	60	36.8	...
4	82	130	30	187	61	36.4	...
5	84	130	28	165	62	36.9	...
6	86	132	35	164	70	37.0	...
...	..	..	..	..	..	..	...

compared to person number three?

PCA

Person	DBP	SBP	BMI	Chol.	Pulse	Temp	...
1	78	126	25	170	55	37.4	...
2	80	128	27	177	56	37.8	...
3	81	127	23	183	60	36.8	...
4	82	130	30	187	61	36.4	...
5	84	130	28	165	62	36.9	...
6	86	132	35	164	70	37.0	...
...	..	..	..	..	..	..	...

When we have many variables, it is very difficult to manually identify persons with a similar health profile.

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

If we combine all the variables into just two variables, that are called PC1 and PC2, it will be a lot easier to identify two individuals who have a similar health profile since we then only need to study two variables instead of many variables.

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

When we combine variables with PCA, we will get these kinds of scores that are centered around zero, which explains why about half of the values are negative.

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

If we combine all the variables into just two, we can see that these two persons have a similar health profile, because they have similar principal component scores,

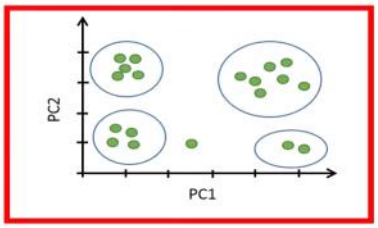
PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

whereas these two persons also have a similar health profile, but different from person number one and two.

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



If we plot these principal component scores in a two dimensional plot like this, each point will represent the combined healthy profile of each individual.

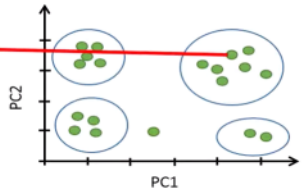
PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

For example, this point might represent the scores of PC1 and PC2, which is the combined health profile for person number one,

PCA

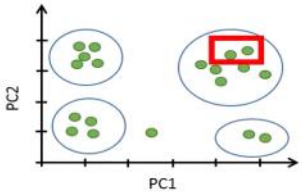
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



whereas this point might represent the health profile for person number two.

PCA

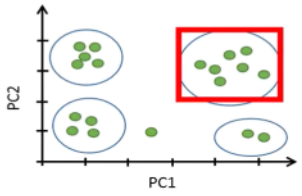
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



Since these two individuals are close in this plot, this suggests that these two individuals have a similar health profile relative to the other individuals.

PCA

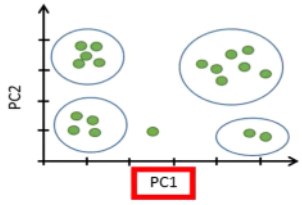
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



We have actually identified seven individuals that seem to have a distinct health profile compared to the other individuals.

PCA

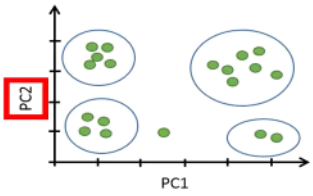
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



Suppose that PC1 is mainly associated with clinical variables associated with the health,

PCA

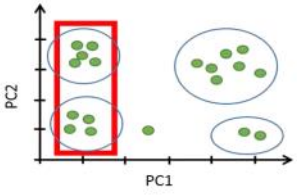
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



whereas PC2 is associated with variables that generally differ between men and women.

PCA

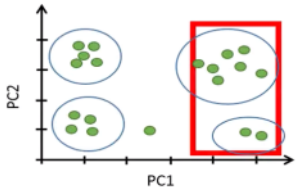
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



Since these individuals have about the same values of the variables associated with the health, they would be considered as having a similar health profile,

PCA

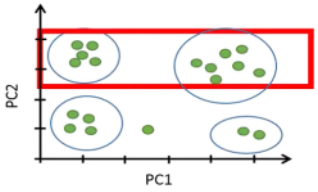
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



whereas these individuals seem to have a different health profile.

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

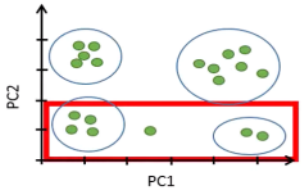


Since PC2 is associated with variables that differ mainly between men and women, these points might therefore represent men,

PCA : the basics - simply explained

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..

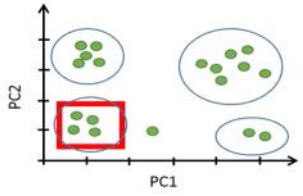


whereas these points might represent women.

17:28 / 22:10

PCA

Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



These four points might therefore represent four women with a similar health profile. In another video, we will see how we can associate certain variables with a certain principal component.

PCA

Person	Gene 1	Gene 2	Gene 3	Gene 4	.....
1	69	1	0	233	
2	70	2	0	231	
3	80	1	0	310	
4	70	0	0	288	
5	50	1	0	277	
6	60	2	0	235	

Another example where PCA is commonly used in biology is when we have measured the gene expression of all our genes, by for example RNA-seq.

PCA

Person	Gene 1	Gene 2	Gene 3	Gene 4	.....
1	69	1	0	233	
2	70	2	0	231	
3	80	1	0	310	
4	70	0	0	288	
5	50	1	0	277	
6	60	2	0	235	

In such case, we might have information on the gene expression of thousands of genes from a certain cell type extracted from each individual.

PCA

Person	Gene 1	Gene 2	Gene 3	Gene 4	.....
1	69	1	0	233	
2	70	2	0	231	
3	80	1	0	310	
4	70	0	0	288	
5	50	1	0	277	
6	60	2	0	235	

For example, in this case we have information of the expression of thousands of genes for each of the six persons.

PCA

Person	Gene 1	Gene 2	Gene 3	Gene 4	.....
1	69	1	0	233	
2	70	2	0	231	
3	80	1	0	310	
4	70	0	0	288	
5	50	1	0	277	
6	60	2	0	235	

For example, we see that gene number three is not expressed in none of the samples from the six individuals,

PCA

Person	Gene 1	Gene 2	Gene 3	Gene 4	.....
1	69	1	0	233	
2	70	2	0	231	
3	80	1	0	310	
4	70	0	0	288	
5	50	1	0	277	
6	60	2	0	235	

whereas gene number four is highly expressed.

PCA

Person	Gene 1	Gene 2	Gene 3	Gene 4	.....
1	69	1	0	233	
2	70	2	0	231	
3	80	1	0	310	
4	70	0	0	288	
5	50	1	0	277	
6	60	2	0	235	

Since we have thousands of genes, it would be very difficult to identify two persons with a similar gene expression. In other words, it will be difficult to identify two persons with a similar gene expression profile.

PCA

Person	PC1	PC2
1	8	2
2	9	3
3	-5	11
4	-4	10
5	-5	-12
6	-3	-14

However, if we combine the expression of all genes into just two variables, PC1 and PC2, it will become a lot easier to identify persons with a similar gene expression profile.

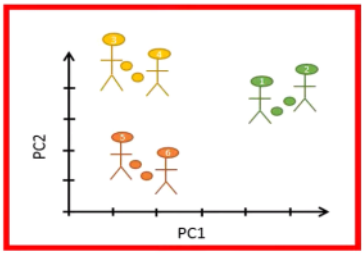
PCA

Person	PC1	PC2
1	8	2
2	9	3
3	-5	11
4	-4	10
5	-5	-12
6	-3	-14

For example, we see that the principal component scores for person number one and two are similar, which suggest that these two persons have a similar gene expression profile.

PCA

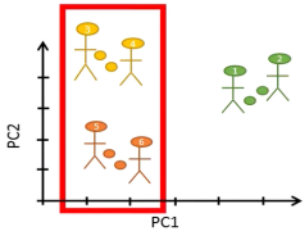
Person	PC1	PC2
1	8	2
2	9	3
3	-5	11
4	-4	10
5	-5	-12
6	-3	-14



If we plot these scores in a two dimensional plot, it becomes even easier to identify two individuals with similar scores.

PCA

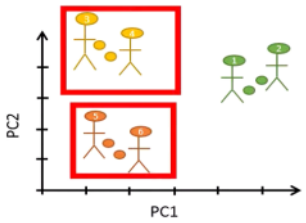
Person	PC1	PC2
1	8	2
2	9	3
3	-5	11
4	-4	10
5	-5	-12
6	-3	-14



These four persons have similar scores for PC1,

PCA

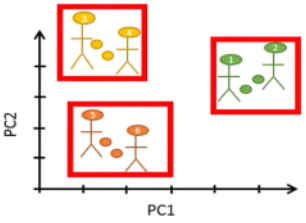
Person	PC1	PC2
1	8	2
2	9	3
3	-5	11
4	-4	10
5	-5	-12
6	-3	-14



but they are separated if we study PC2.

PCA

Person	PC1	PC2
1	8	2
2	9	3
3	-5	11
4	-4	10
5	-5	-12
6	-3	-14



We can identify these pairs of individuals as having a similar gene expression profiles.

PCA

Cell	Gene 1	Gene 2	Gene 3	Gene 4	....
Cell 1	0	20	5	0	....
Cell 2	10	0	56	0	....
Cell 3	15	0	20	0	....
Cell 4	10	13	13	0	....
Cell 5	42	55	60	0	....
Cell 6	0	0	30	0	....
...	..	...	...	...	

Similarly, we might have data from single-cell RNA-seq, where the gene expression of each extracted cell has been measured from just one person.

PCA

Cell	Gene 1	Gene 2	Gene 3	Gene 4	....
Cell 1	0	20	5	0	....
Cell 2	10	0	56	0	....
Cell 3	15	0	20	0	....
Cell 4	10	13	13	0	....
Cell 5	42	55	60	0	....
Cell 6	0	0	30	0	....
...	..	...	...	...	

For example, this row shows the expression of all the genes in cell number one,

PCA

Cell	Gene 1	Gene 2	Gene 3	Gene 4	....
Cell 1	0	20	5	0	....
Cell 2	10	0	56	0	....
Cell 3	15	0	20	0	....
Cell 4	10	13	13	0	....
Cell 5	42	55	60	0	....
Cell 6	0	0	30	0	....
...	..	...	...	...	

whereas this row shows the expression of all the genes in cell number two.

PCA

Cell	Gene 1	Gene 2	Gene 3	Gene 4	....
Cell 1	0	20	5	0	....
Cell 2	10	0	56	0	....
Cell 3	15	0	20	0	....
Cell 4	10	13	13	0	....
Cell 5	42	55	60	0	....
Cell 6	0	0	30	0	....
...	..	...	...	...	

This column shows the gene expression of gene number 4 for each cell. As can be seen, this gene is not expressed in the first six cells.

PCA

Cell	Gene 1	Gene 2	Gene 3	Gene 4	....
Cell 1	0	20	5	0	....
Cell 2	10	0	56	0	....
Cell 3	15	0	20	0	....
Cell 4	10	13	13	0	....
Cell 5	42	55	60	0	....
Cell 6	0	0	30	0	....
...	..	...	...	...	

The aim with these types of studies is usually to identify cells that have a similar gene expression as such cells might represent a unique cell type or cells that undergo a certain biological process.

PCA

Cell	Gene 1	Gene 2	Gene 3	Gene 4	....
Cell 1	0	20	5	0	....
Cell 2	10	0	56	0	....
Cell 3	15	0	20	0	....
Cell 4	10	13	13	0	....
Cell 5	42	55	60	0	....
Cell 6	0	0	30	0	....
...	..	...	...	...	

However, this kind of data set has thousands of columns and may have thousands of rows. To manually identify a group of cells that have a similar gene expression profile would be almost impossible.

PCA

Cell	PC1	PC2
Cell 1	5	22
Cell 2	-10	0
Cell 3	-3	13
Cell 4	-3	13
Cell 5	19	55
Cell 6	-7	-7
...	..	...

However, if we could reduce the number of columns to just two, it will be a lot easier to identify cells with a similar gene expression profile.

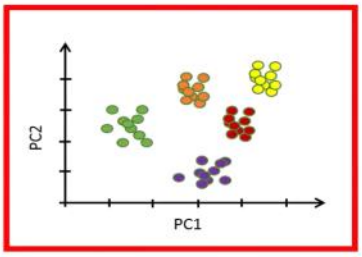
PCA

Cell	PC1	PC2
Cell 1	5	22
Cell 2	-10	0
Cell 3	-3	13
Cell 4	-3	13
Cell 5	19	55
Cell 6	-7	-7
...	..	...

For example, cell number three and four seem to have similar gene expressions since they have similar scores.

PCA

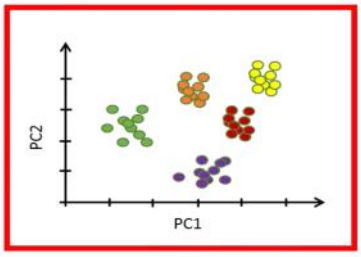
Cell	PC1	PC2
Cell 1	5	22
Cell 2	-10	0
Cell 3	-3	13
Cell 4	-3	13
Cell 5	19	55
Cell 6	-7	-7
...	..	...



If we plot these two principal components in a two-dimensional plot, it will be quite easy to identify cells with similar gene expressions.

PCA

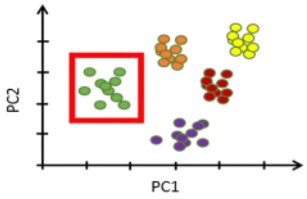
Cell	PC1	PC2
Cell 1	5	22
Cell 2	-10	0
Cell 3	-3	13
Cell 4	-3	13
Cell 5	19	55
Cell 6	-7	-7
...	..	...



Each point in this plot represents the combined gene expression of a single cell. In this plot, we have about 50 points, which represent 50 cells.

PCA

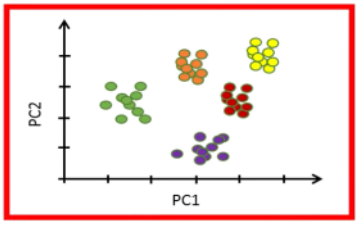
Cell	PC1	PC2
Cell 1	5	22
Cell 2	-10	0
Cell 3	-3	13
Cell 4	-3	13
Cell 5	19	55
Cell 6	-7	-7
...	..	...



Cells that can be found in a distinct cluster can be seen as cells having similar gene expressions across all genes.

PCA

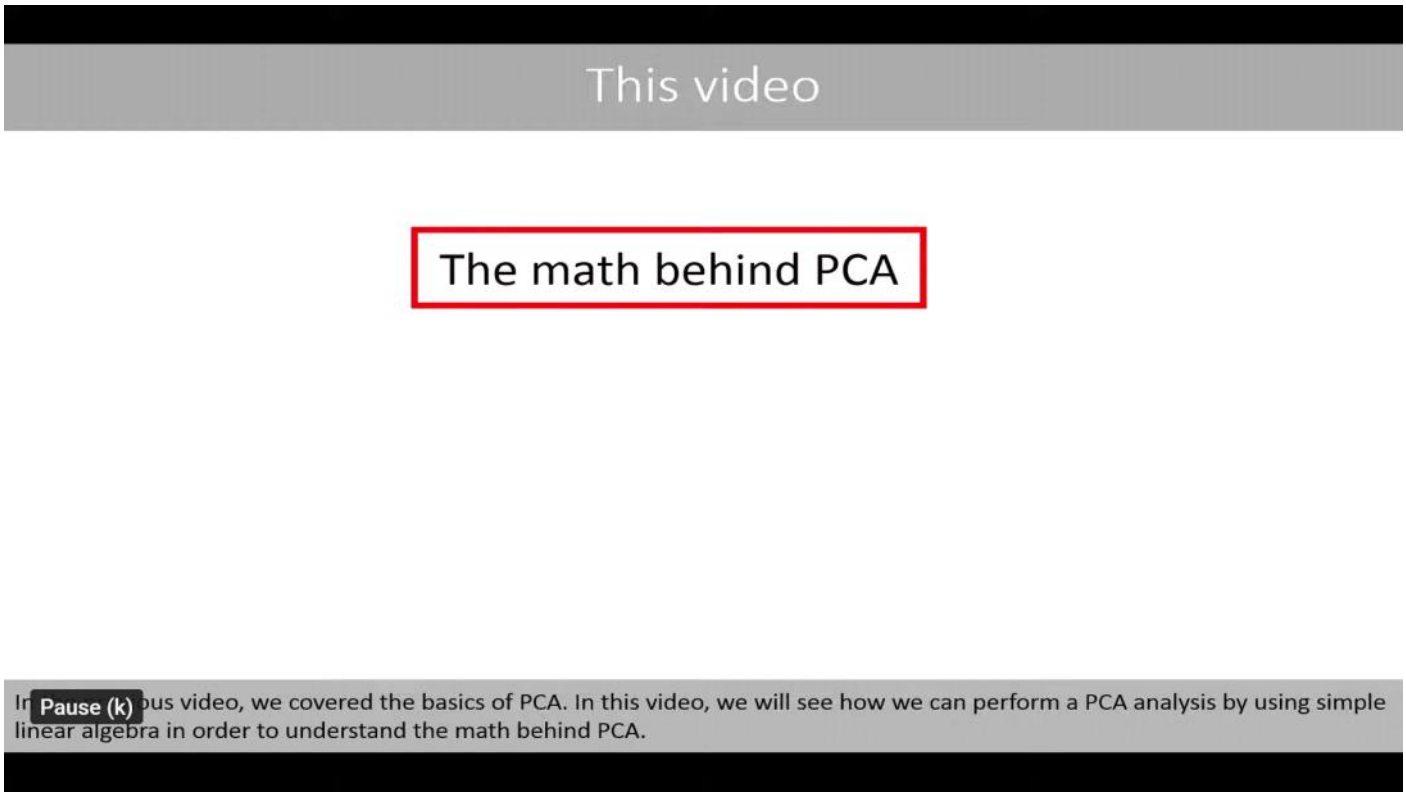
Cell	PC1	PC2
Cell 1	5	22
Cell 2	-10	0
Cell 3	-3	13
Cell 4	-3	13
Cell 5	19	55
Cell 6	-7	-7
...	..	...

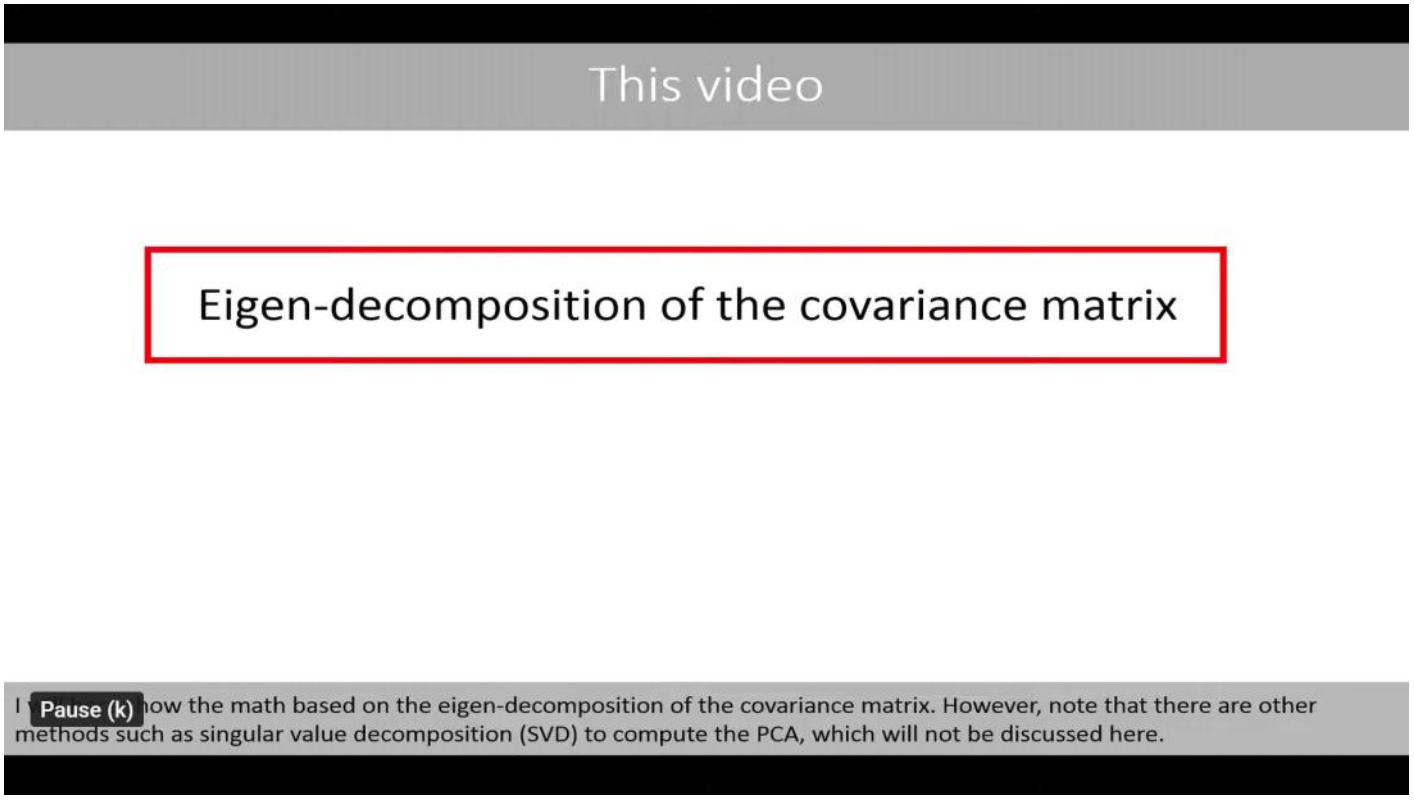


With this kind of plot, it is possible to identify different cell types and cells that undergo a certain biological process.

PCA MATH

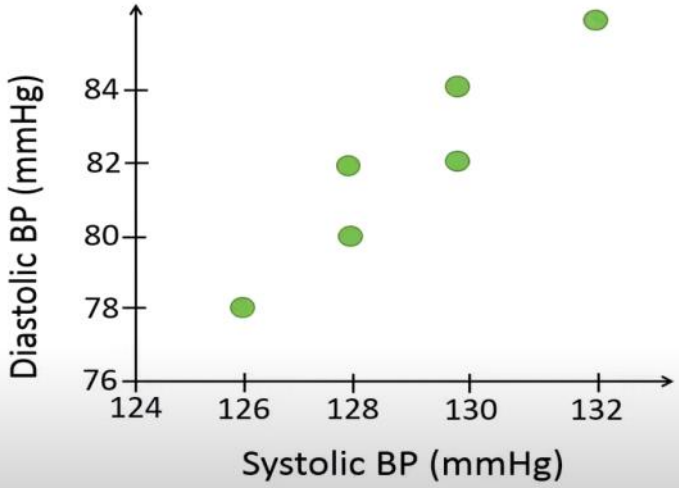
10 October 2023 10:02





PCA : the math - step-by-step with a simple example

Example data

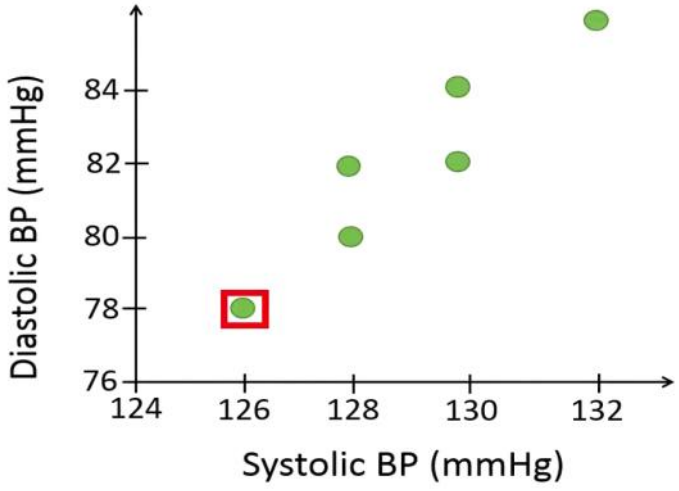


Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

To explain how the PCA works, we will use the following example data. We will use PCA to combine the two blood pressure variables into just one variable based on data from six individuals.

0:39 / 20:21 • Example data >

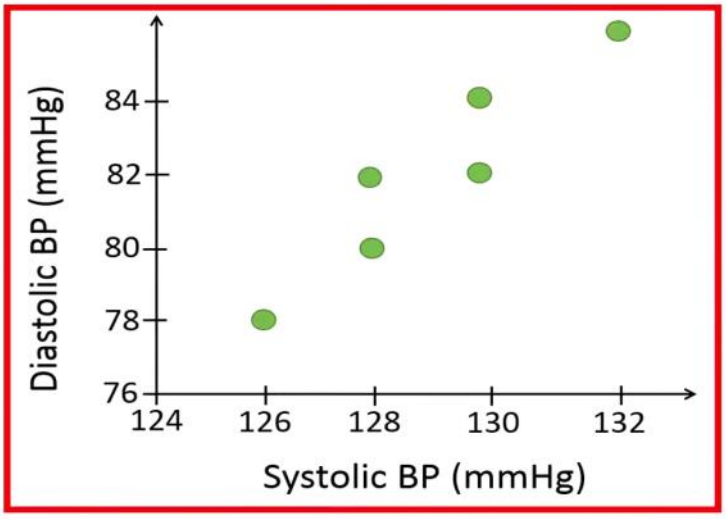
Example data



Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

For example, person number one has a diastolic blood pressure of 78 and a systolic blood pressure of 126,

Example data



Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

For this data set, it seems to be a strong positive correlation between the two variables.

PCA

1. Center the data
2. Calculate the covariance matrix
3. Calculate eigenvalues of the covariance matrix
4. Calculate eigenvectors of the covariance matrix
5. Order the eigenvectors
6. Calculate the principal components

To compute a PCA, we can perform the following steps,

PCA

1. Center the data
2. Calculate the covariance matrix
3. Calculate eigenvalues of the covariance matrix
4. Calculate eigenvectors of the covariance matrix
5. Order the eigenvectors
6. Calculate the principal components

where we first center our data.

PCA

1. Center the data
2. Calculate the covariance matrix
3. Calculate eigenvalues of the covariance matrix
4. Calculate eigenvectors of the covariance matrix
5. Order the eigenvectors
6. Calculate the principal components

Then, we calculate the covariance matrix on our centered data.

PCA

1. Center the data
2. Calculate the covariance matrix
3. Calculate eigenvalues of the covariance matrix
4. Calculate eigenvectors of the covariance matrix
5. Order the eigenvectors
6. Calculate the principal components

Next, we compute the eigenvalues and eigenvectors of the covariance matrix.

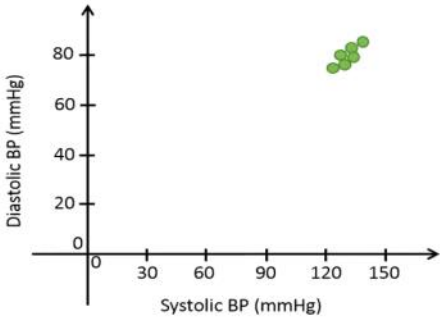
PCA

1. Center the data
2. Calculate the covariance matrix
3. Calculate eigenvalues of the covariance matrix
4. Calculate eigenvectors of the covariance matrix
5. Order the eigenvectors
6. Calculate the principal components

Finally, we order the eigenvectors and calculate the principal components, which is our combined or transformed data set.

1. Center the data

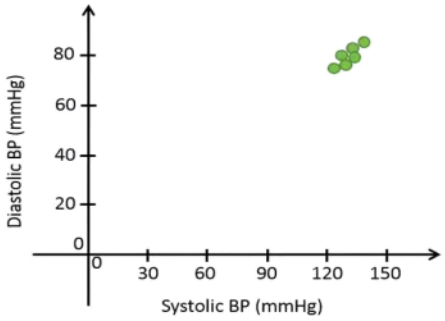
Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86



Usually, one starts to center or standardize the data in the first step of the PCA analysis. In this case, we will only center the data, which means that we subtract all the values for each variable by its corresponding mean.

1. Center the data

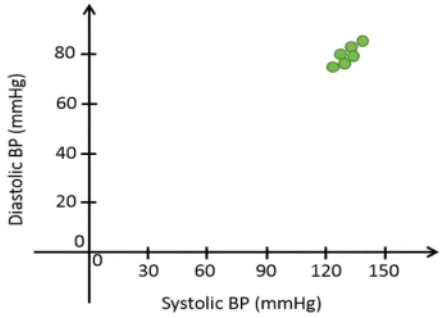
Systolic BP	Diastolic BP
126 -129 = -3	78
128 -129 = -1	80
128 -129 = -1	82
130 -129 = 1	82
130 -129 = 1	84
132 -129 = 3	86



We therefore subtract the mean systolic blood pressure,

1. Center the data

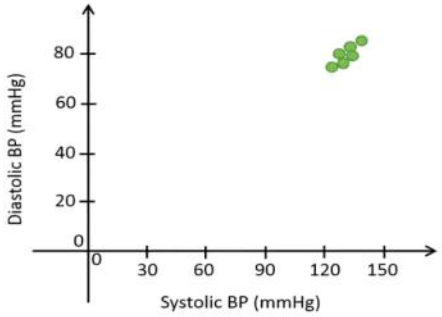
Systolic BP	Diastolic BP
126 -129 = -3	78
128 -129 = -1	80
128 -129 = -1	82
130 -129 = 1	82
130 -129 = 1	84
132 -129 = 3	86



from the individual observations.

1. Center the data

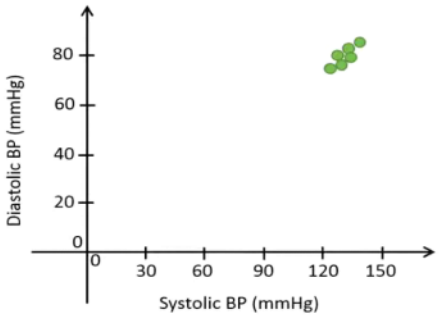
Systolic BP	Diastolic BP
126 -129 = -3	78
128 -129 = -1	80
128 -129 = -1	82
130 -129 = 1	82
130 -129 = 1	84
132 -129 = 3	86



Centering the systolic blood pressure results in the following values, which tell how far away the original values are from the mean.

1. Center the data

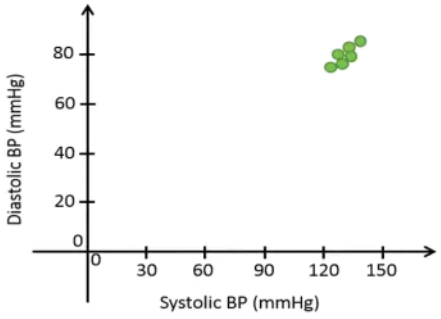
Systolic BP	Diastolic BP
126 -129 = -3	78 -82 = -4
128 -129 = -1	80 -82 = -2
128 -129 = -1	82 -82 = 0
130 -129 = 1	82 -82 = 0
130 -129 = 1	84 -82 = 2
132 -129 = 3	86 -82 = 4



We then do the same calculations for the diastolic blood pressure, which has a mean value of 82.

1. Center the data

Systolic BP	Diastolic BP
126 -129 = -3	78 -82 = -4
128 -129 = -1	80 -82 = -2
128 -129 = -1	82 -82 = 0
130 -129 = 1	82 -82 = 0
130 -129 = 1	84 -82 = 2
132 -129 = 3	86 -82 = 4



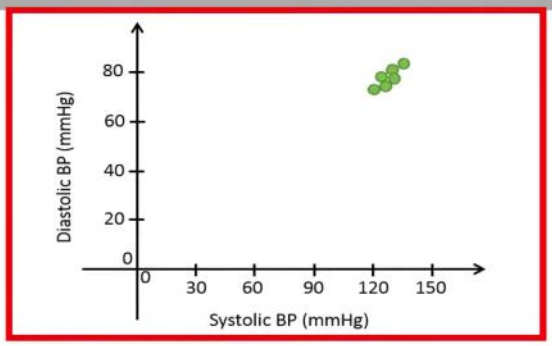
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

We can summarize the centered data in the following table.

1. Center the data

Systolic BP	Diastolic BP
126 -129 = -3	78 -82 = -4
128 -129 = -1	80 -82 = -2
128 -129 = -1	82 -82 = 0
130 -129 = 1	82 -82 = 0
130 -129 = 1	84 -82 = 2
132 -129 = 3	86 -82 = 4

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

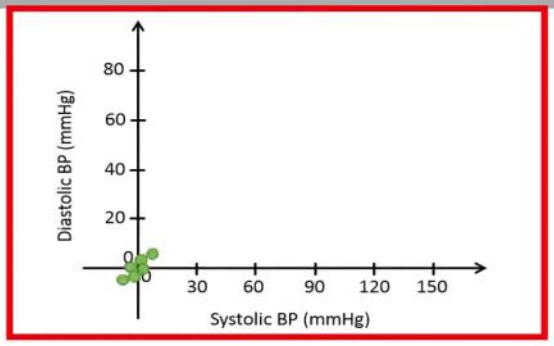


When we center the data, it means that we center the data points around the origin. Centering the data around the origin will help us later when we will rotate the data.

1. Center the data

Systolic BP	Diastolic BP
126 -129 = -3	78 -82 = -4
128 -129 = -1	80 -82 = -2
128 -129 = -1	82 -82 = 0
130 -129 = 1	82 -82 = 0
130 -129 = 1	84 -82 = 2
132 -129 = 3	86 -82 = 4

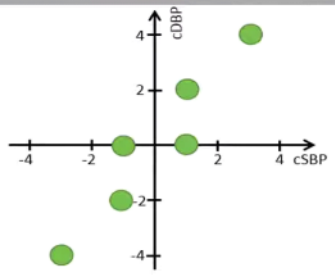
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



When we center the data, it means that we center the data points around the origin. Centering the data around the origin will help us later when we will rotate the data.

1. Center the data

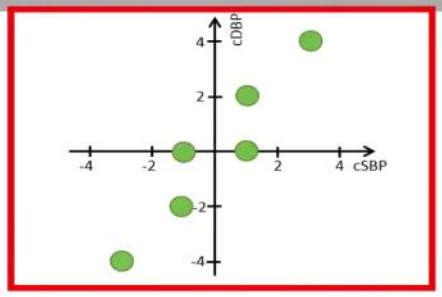
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



After we have centered the data, we will have the following values,

1. Center the data

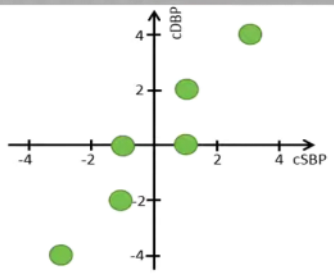
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



which can be plotted like this, where the x-axis now represents the centered systolic blood pressure, whereas the y-axis represents the centered diastolic blood pressure.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

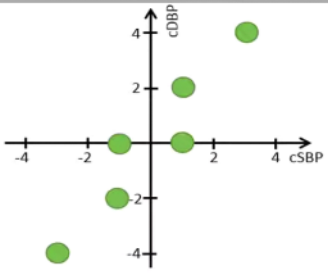


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Next, we calculate the covariance matrix based on the centered data. Note that we would have got the same values in the covariance matrix if we instead would have used the original data since the variance does not change when we center the data.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

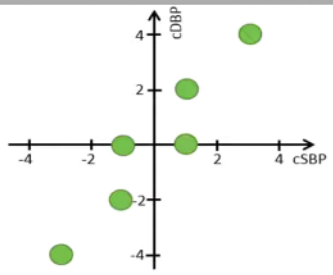


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Remember that the main diagonal of the covariance matrix includes the variance of each variable.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

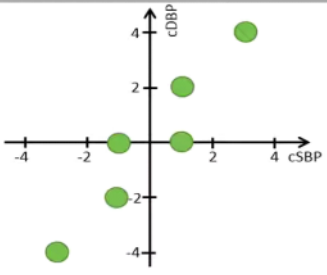


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Next, we calculate the covariance matrix based on the centered data. Note that we would have got the same values in the covariance matrix if we instead would have used the original data since the variance does not change when we center the data.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

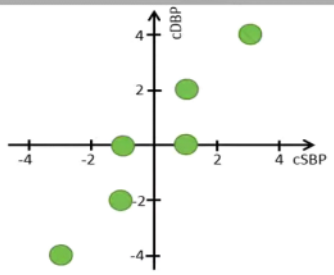


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Remember that the main diagonal of the covariance matrix includes the variance of each variable.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



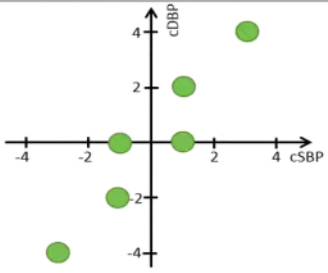
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var(cSBP)} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

The sample variance of the centered systolic blood pressure is calculated like this.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



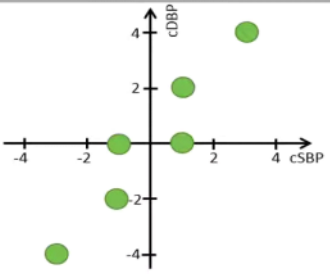
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var(cSBP)} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

When we calculate the variance of the centered data, the calculations become a bit simpler since the mean of the centered data is always equal to zero.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



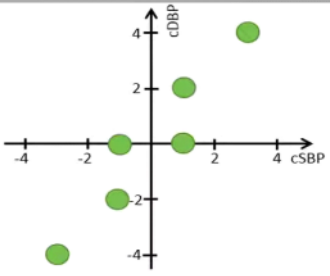
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2) / (6-1) = 22/5 = 4.4$$

To calculate the sample variance of the centered data, we therefore simply sum the squared values,

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



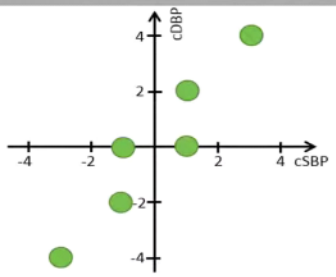
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2) / (6-1) = 22/5 = 4.4$$

and divide by the sample size minus one.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

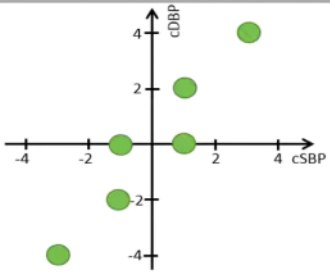
$$\text{var(cSBP)} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

$$\text{var(cDBP)} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

Then we calculate the variance of the diastolic blood pressure by using the same equation.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var(cSBP)} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

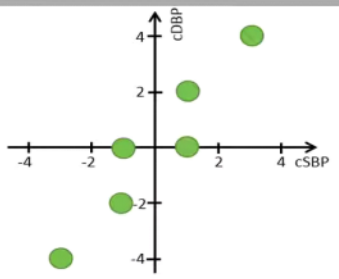
$$\text{var(cDBP)} = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov(cSBP,cDBP)} = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = ((-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4)/(6-1) = 28/5 = 5.6$$

Finally, we calculate the covariance, which is a measure of how much the two variables spread together.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

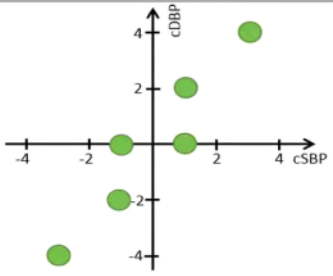
$$\text{var}(\text{cDBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(\text{cSBP},\text{cDBP}) = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = (-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4)/(6-1) = 28/5 = 5.6$$

The sample covariance is calculated by multiplying the centered values of the two variables.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

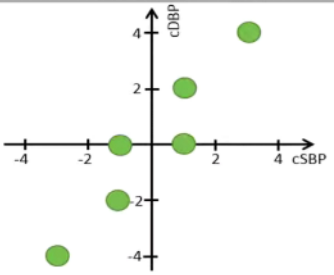
$$\text{var}(\text{cDBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(\text{cSBP},\text{cDBP}) = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = (-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4)/(6-1) = 28/5 = 5.6$$

For example, we multiply the centered values for person number one,

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

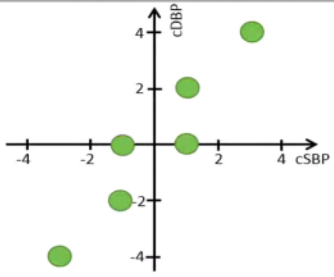
$$\text{var}(\text{cDBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(\text{cSBP},\text{cDBP}) = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = ((-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4)/(6-1) = 28/5 = 5.6$$

and add that to the product of the centered values for person number two, and so on.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

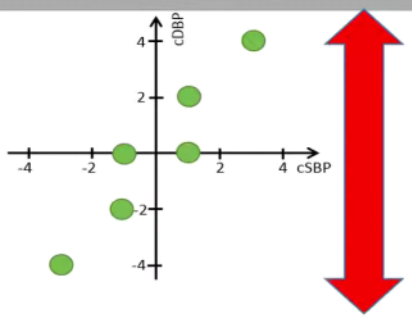
$$\text{var}(\text{cDBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(\text{cSBP},\text{cDBP}) = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = ((-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4)/(6-1) = 28/5 = 5.6$$

Finally, we divide the sum of the products by the sample size minus one.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

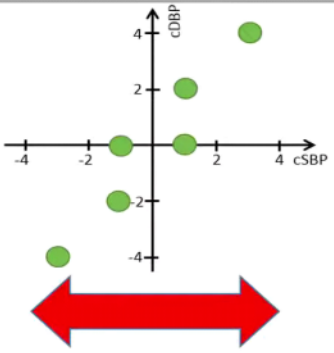


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

We see that the spread in the diastolic blood pressure is a bit higher compared to

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

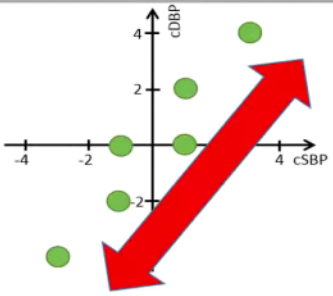


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

the spread in the systolic blood pressure.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

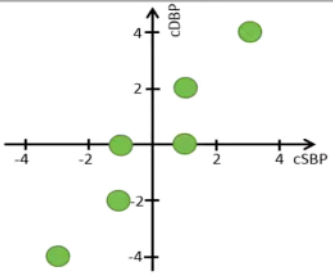


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

The covariance is somewhere between these two values.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



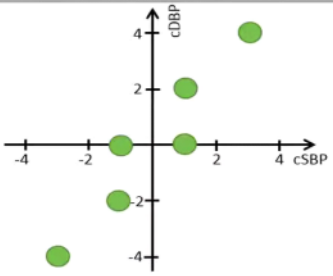
$$\det|A - \lambda I| = 0$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Next, we calculate the eigenvalues of the covariance matrix. For more details, watch the lecture about eigenvalues and eigenvectors.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A - \lambda I| = 0$$

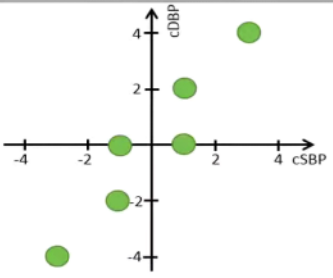
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left|\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = 0$$

We substitute A by the covariance matrix,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A - \lambda I| = 0$$

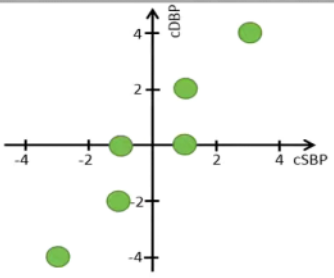
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left|\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = 0$$

and this term by lambda times the identity matrix, which has the same number of rows and columns as the covariance matrix.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det[A - \lambda I] = 0$$

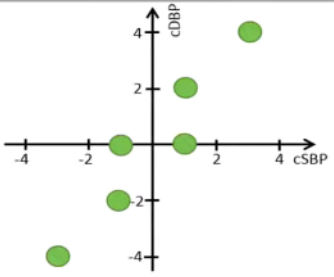
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left[\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right] = 0$$

Subtracting these two matrices results in the following matrix.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A - \lambda I| = 0$$

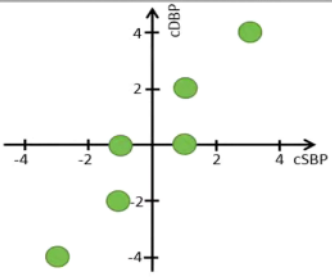
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left|\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right| = 0$$

Next, we calculate the determinant of this matrix,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A-\lambda I|=0$$

$$(4.4-\lambda)(8.0-\lambda)-5.6\cdot 5.6=0$$

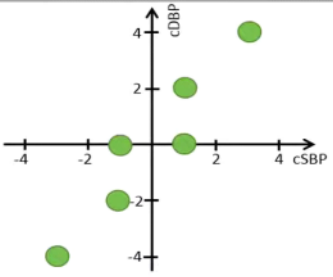
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left[\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right]=0$$

which is the product of this diagonal,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A-\lambda I|=0$$

$$(4.4-\lambda)(8.0-\lambda)-5.6\cdot 5.6=0$$

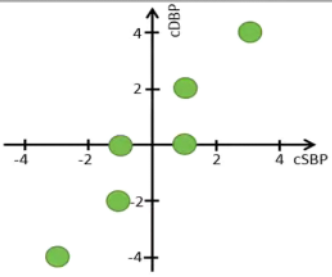
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left[\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right]=0$$

minus,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A-\lambda I|=0$$

$$(4.4-\lambda)(8.0-\lambda)-5.6\cdot 5.6=0$$

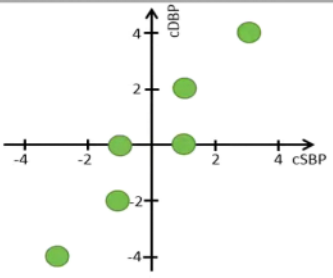
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left|\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right|=0$$

the product of this diagonal.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A-\lambda I|=0$$

$$(4.4-\lambda)(8.0-\lambda)-5.6\cdot 5.6=0$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

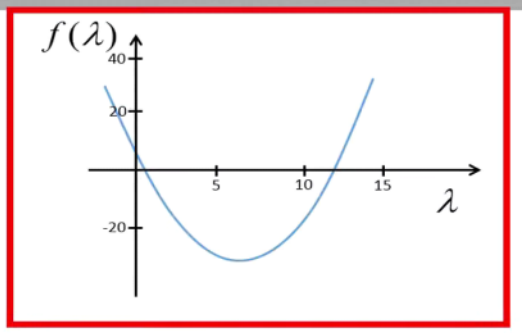
$$\det\left|\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right|=0$$

$$3.84-12.4\lambda+\lambda^2=0$$

After some simplifications, we have the following quadratic equation. Quadratic equations like this can be solved in different ways, which will not be discussed here.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



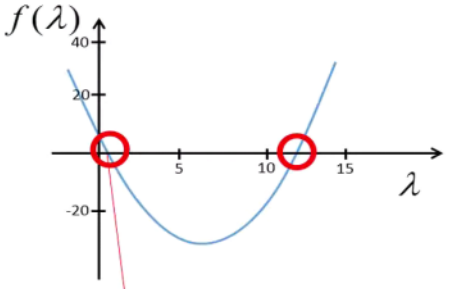
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$3.84 - 12.4\lambda + \lambda^2 = 0$$

However, if we plot how the left-hand side changes as a function of different values of lambda, we see that the left-hand side is equal to zero when lambda is equal to either,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

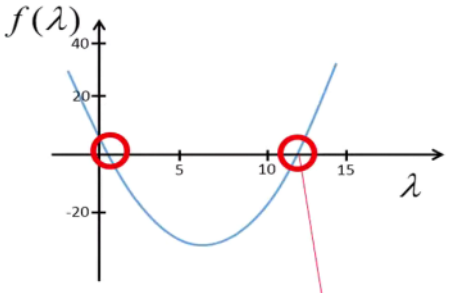
$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\lambda_1 = 0.32 \qquad \lambda_2 = 12.08$$

about 0.32,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

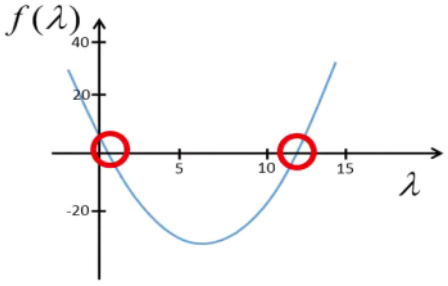
$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\lambda_1 = 0.32 \qquad \lambda_2 = 12.08$$

or 12.08.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

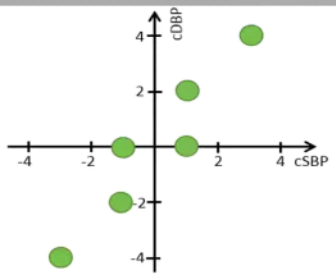
$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\lambda_1 = 0.32 \qquad \lambda_2 = 12.08$$

This means that if we set lambda to either 0.32 or 12.08, the left-hand side of this equation will become equal to zero, or close to zero due to rounding effects in this example.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



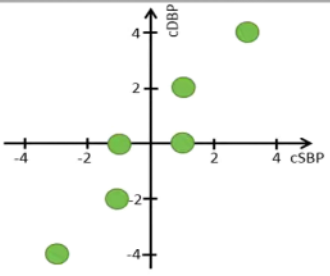
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$\lambda_1 = 0.32 \qquad \lambda_2 = 12.08$

These two values represent our eigenvalues of the covariance matrix.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

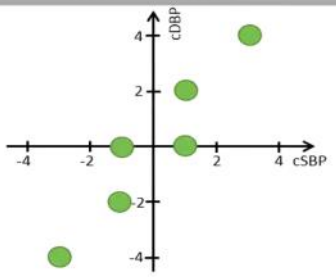
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$\lambda_1 = 0.32 \qquad \lambda_2 = 12.08$

Next, we calculate the corresponding eigenvectors to these two eigenvalues. We will start by calculating the eigenvector of the covariance matrix with the corresponding eigenvalue 12.08.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



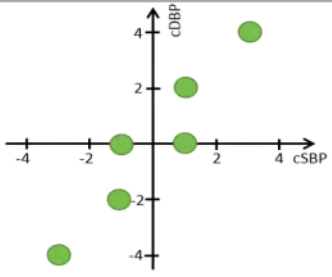
$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

To calculate the eigenvectors of the covariance matrix, we use the following equation that we have discussed in a previous video about eigenvectors.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

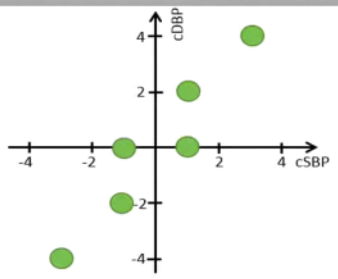
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

We plug in the covariance matrix,

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

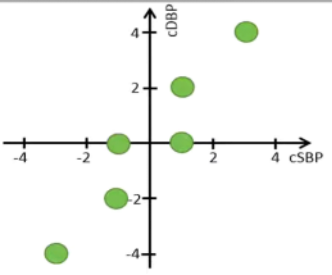
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \begin{bmatrix} x \\ y \end{bmatrix}$$

and one of the two eigenvalues.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

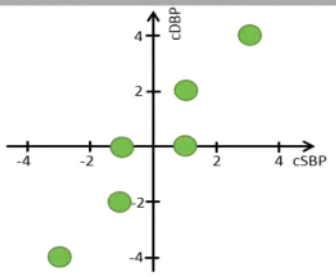
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

If we multiply the covariance matrix by this column vector,

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

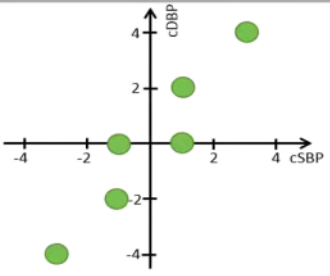
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

and multiply the eigenvalue by the same vector,

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

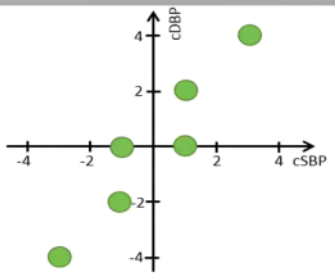
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\begin{aligned} 4.4x + 5.6y &= 12.08x \\ 5.6x + 8.0y &= 12.08y \end{aligned}$$

we will get the following system of equations.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

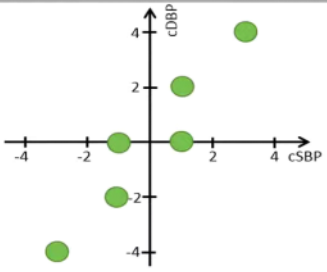
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$

We move these two terms to the right-hand side.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

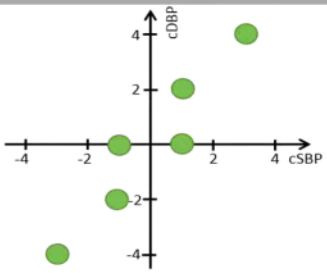
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$
$$5.6y = 7.68x$$
$$5.6x = 4.08y$$

After some simplifications, we have the following system of equations.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

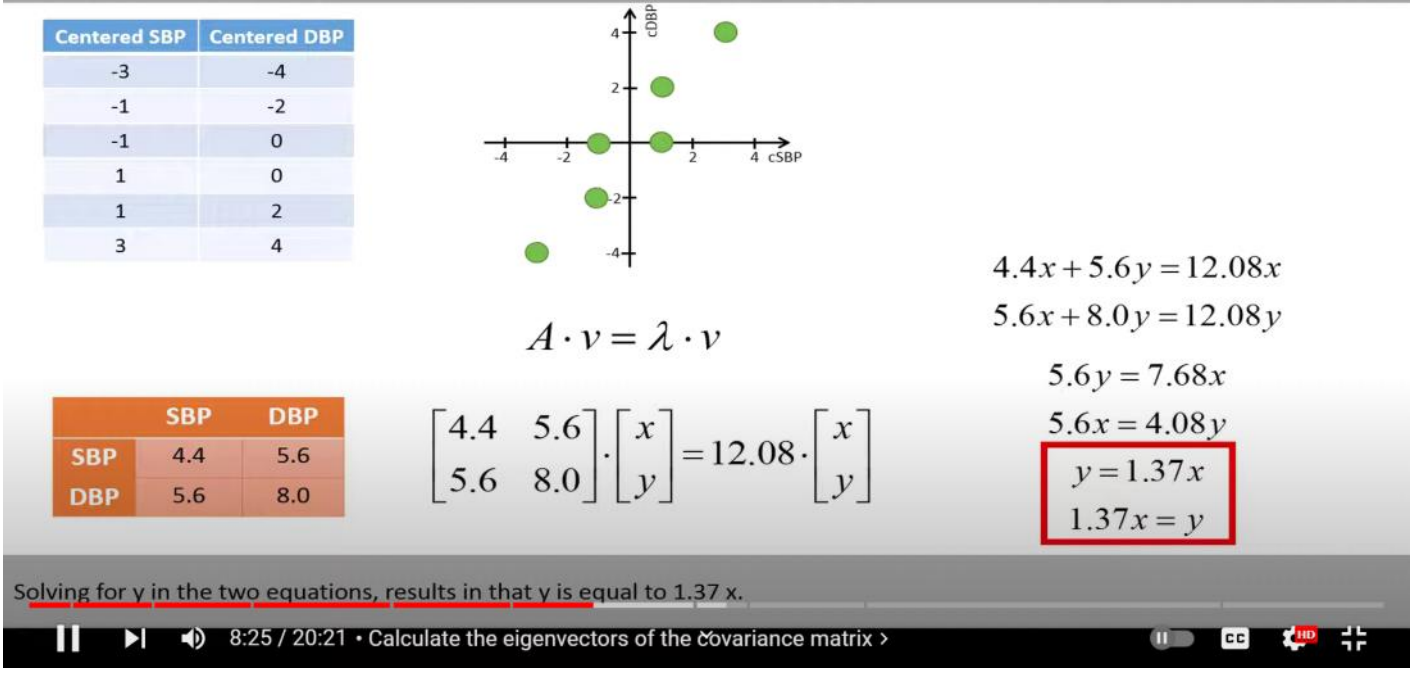
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$
$$5.6y = 7.68x$$
$$5.6x = 4.08y$$
$$y = 1.37x$$
$$1.37x = y$$

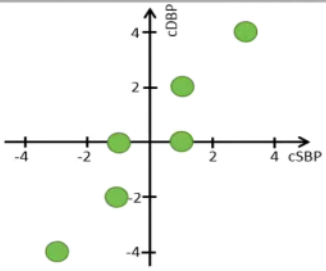
Solving for y in the two equations, results in that y is equal to 1.37 x.

PCA : the math - step-by-step with a simple example
4. Calculate the eigenvectors of the covariance matrix



4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

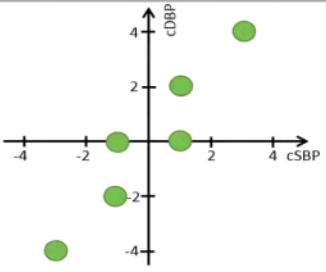
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$
$$5.6y = 7.68x$$
$$5.6x = 4.08y$$
$$y = 1.37x$$
$$1.37x = y$$

For example, if we set x equal to one,

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

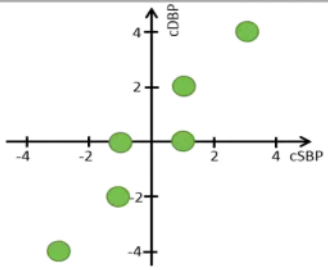
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$
$$5.6y = 7.68x$$
$$5.6x = 4.08y$$
$$y = 1.37x$$
$$1.37x = y$$

y is equal to 1.37.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



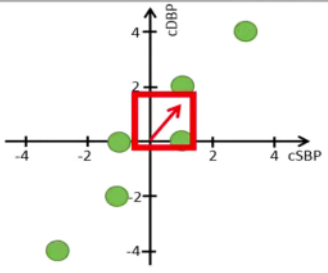
$$v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

This vector is therefore an eigenvector of the covariance matrix.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



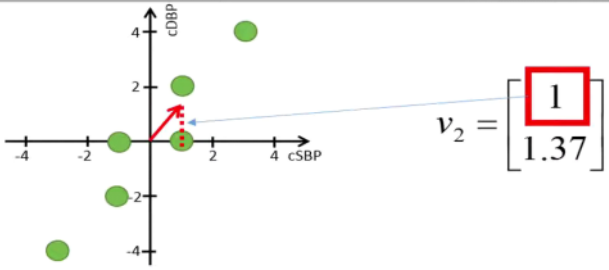
$$v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

We can illustrate this vector in the plot like this, by drawing an arrow from the origin to the coordinates,

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

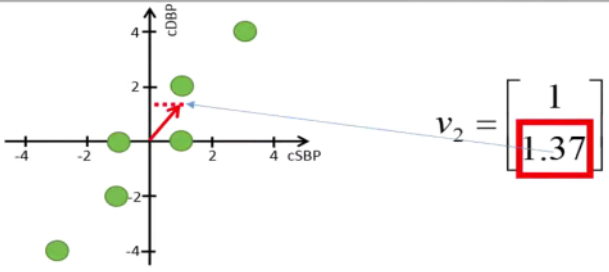


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

one

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

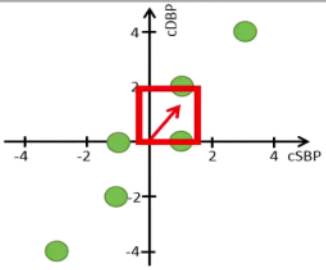


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

and 1.37.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



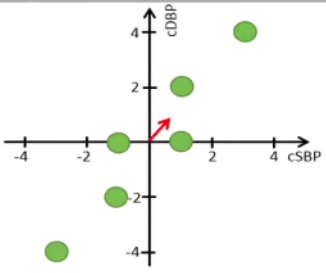
$$v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

We will now normalize this vector to unit length, which means that it should have a length of one. Watch the lecture about the eigenvectors and eigenvalues to see how one can normalize the eigenvector to unit length.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



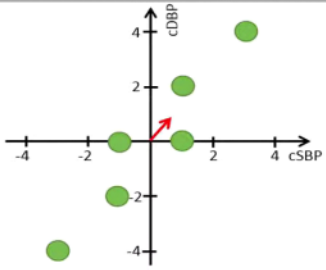
$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

After normalization, this vector represents one out of two eigenvectors of the covariance matrix.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

$$\lambda_1 = 0.32$$

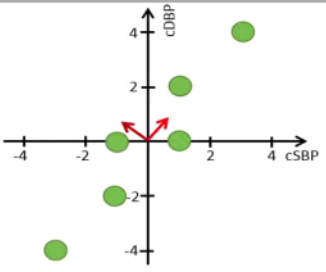
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 0.32 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

To find the second eigenvector, we do the same calculations as before based on the second eigenvalue.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

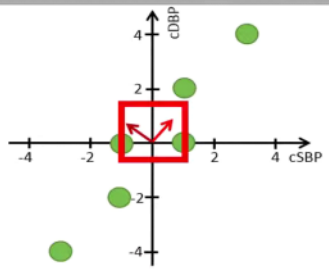
$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

After some calculations, this vector represents our second eigenvector with unit length.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

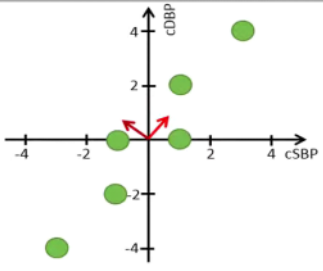
$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Since the covariance matrix is a symmetric matrix, the eigenvectors will be orthogonal, which means that the angle between them is 90 degrees.

5. Order the eigenvectors

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

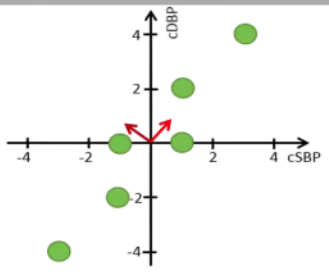
$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Next, we order the eigenvectors based on their corresponding eigenvalues, where the eigenvector with the largest eigenvalue becomes our first eigenvector.

5. Order the eigenvectors

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

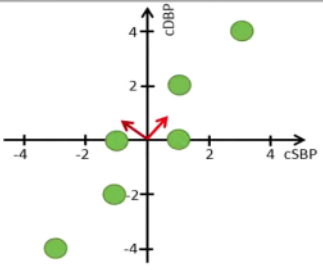
$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Since this eigenvector has the largest eigenvalue, it will represent our first eigenvector. We therefore rename this vector so that it is called v1 instead of v2.

5. Order the eigenvectors

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

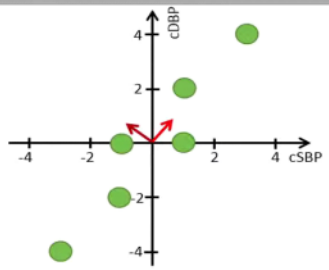
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Let's put these two eigenvectors together into a matrix that we call V,

5. Order the eigenvectors

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

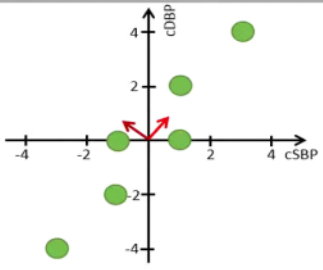
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

where the first column represents the first eigenvector with the highest eigenvalue,

5. Order the eigenvectors

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

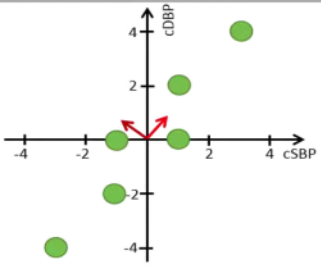
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

and the second column represents our second eigenvector.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

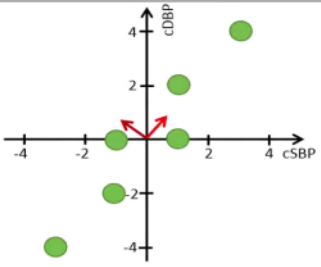


$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

We will now use this matrix to transform our original centered data so that the two variables are completely uncorrelated.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



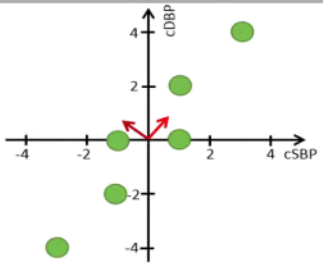
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

$$D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix}$$

We define a matrix D, which includes our centered data.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

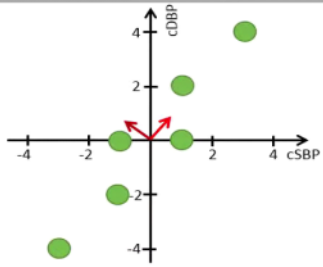


$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

Next, we multiply our data matrix D by matrix V, which includes our eigenvectors as columns.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

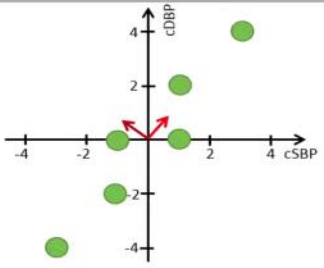


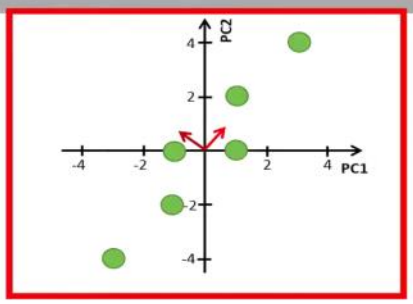
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

Then we get a new matrix with the transformed data. This transformed data is called principal component scores, or just scores, which represent the original centered data in the principal component space.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



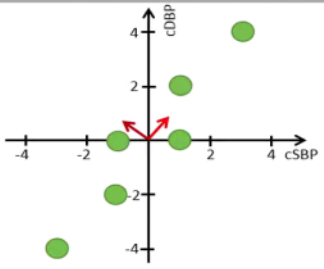


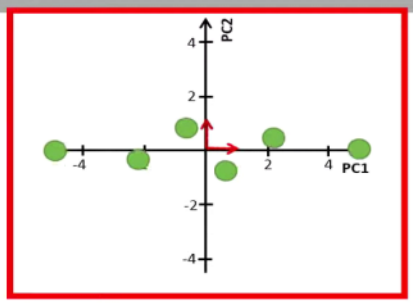
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

When we go from our original data matrix to the transformed data, this can be seen like we rotate the data clockwise until the two eigenvectors point in the same direction as the x and y-axes of the plot.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



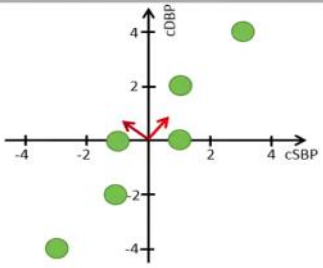


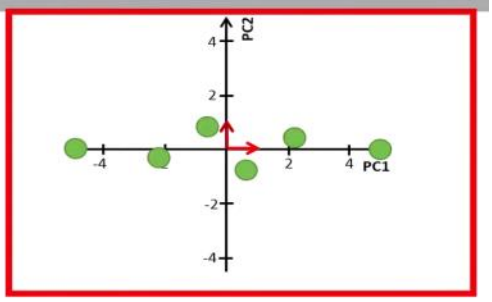
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

When we go from our original data matrix to the transformed data, this can be seen like we rotate the data clockwise until the two eigenvectors point in the same direction as the x and y-axes of the plot.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



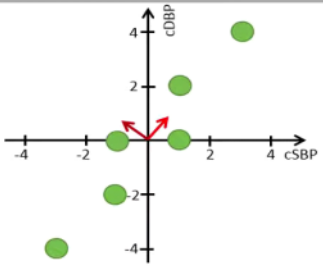


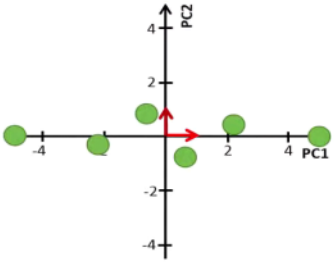
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

The rotated data now looks like this. Note that the labels of the axes have now been changed to principal component one and two.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



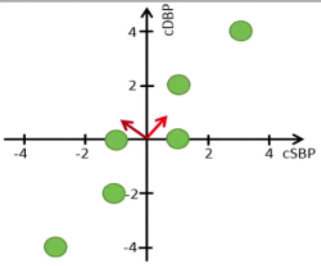


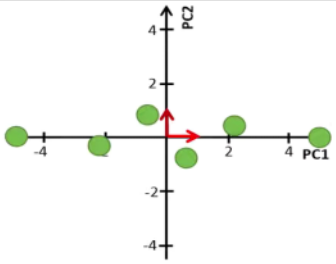
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{array}{c} \begin{matrix} \text{PC1} & \text{PC2} \end{matrix} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{array}$$

Let's call the two columns of the transformed data PC1 and PC2.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



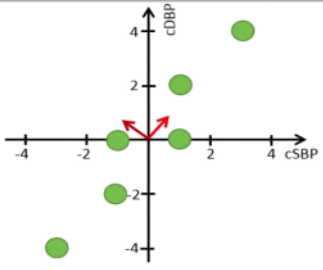


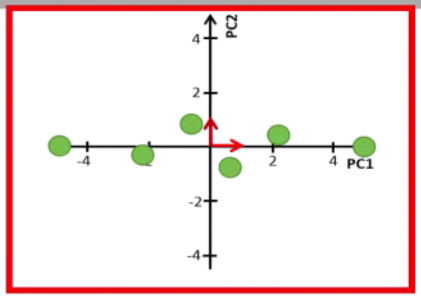
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{matrix} \text{PC1} & \text{PC2} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{matrix}$$

If we would plot this data, where we label the x-axis as PC1 and the y-axis as PC2,

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

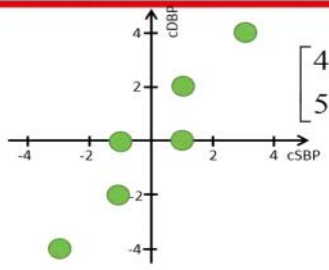




$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{matrix} \text{PC1} & \text{PC2} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{matrix}$$

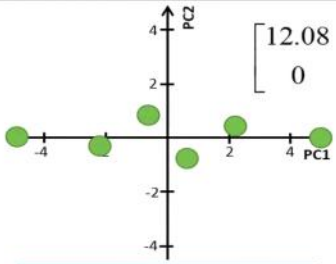
we would get the following plot, which represents the original plot after the rotation. Since we plot the principal component scores, this kind of plot is called a score plot.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

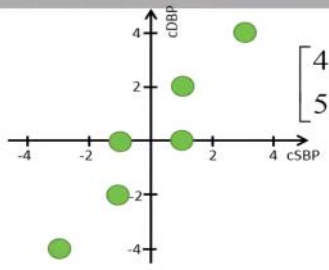


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

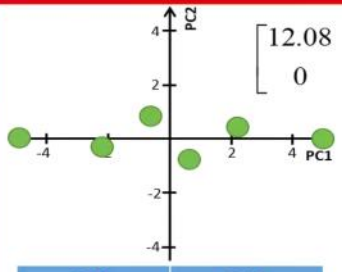
Let's compare the centered data,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

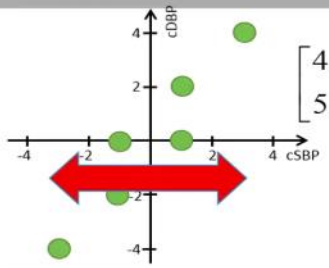


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

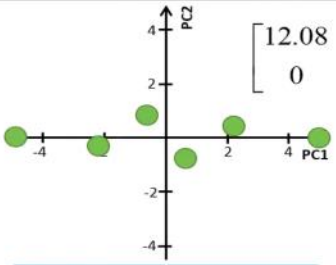
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

with the transformed data.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



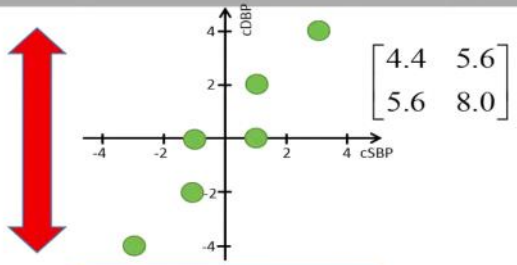
$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

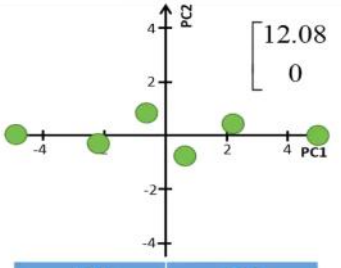
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

The variance of the systolic blood pressure is 4.4,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



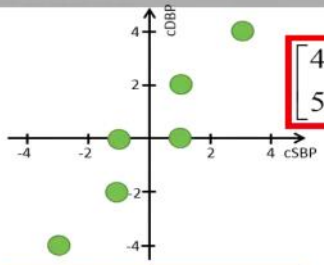
$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

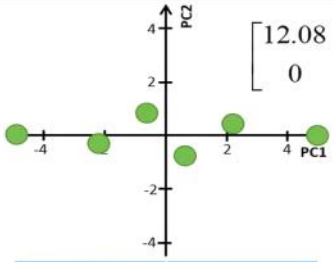
whereas the variance of the diastolic blood pressure is 8.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

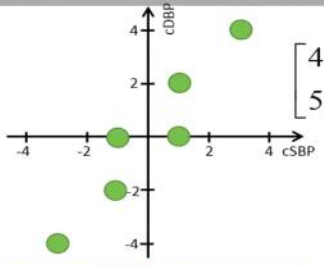


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

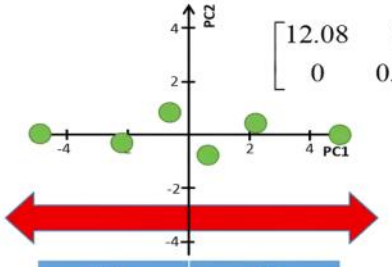
This is the covariance matrix of the data. We see that the covariance is 5.6, which tells us that there is a positive correlation between the two variables.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

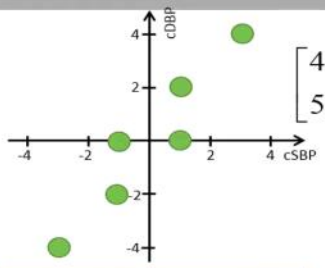


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

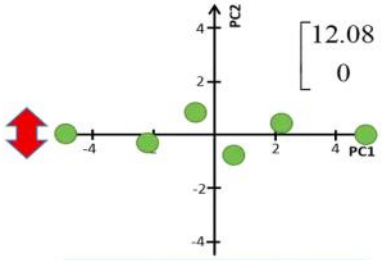
When we transform the data, using PCA, the first variable called PC1 has a variance of 12.08,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

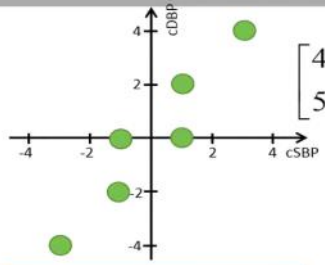


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

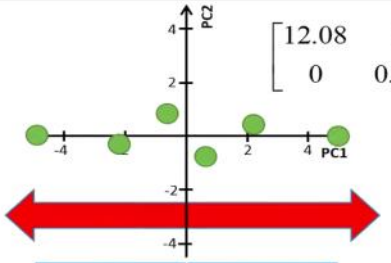
whereas PC2 has only a variance of 0.32.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

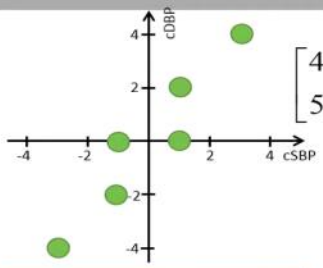


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

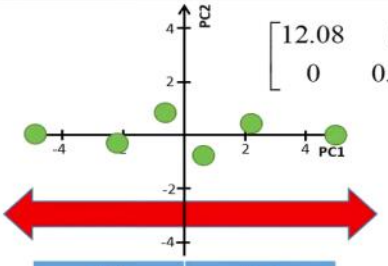
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

This means that almost all variance is kept in the first principal component.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

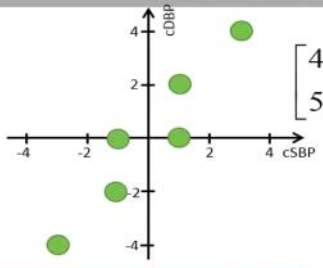
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\%\,\text{var} = \frac{12.08}{12.08+0.32} = 97.4\%$$

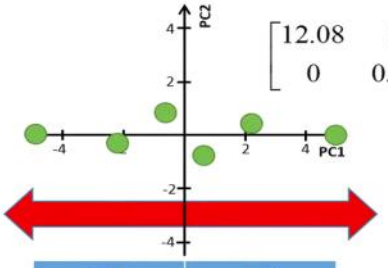
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

If we divide the variance of the first principal component,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

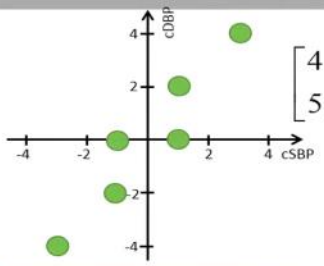
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\%\,\text{var} = \frac{12.08}{12.08+0.32} = 97.4\%$$

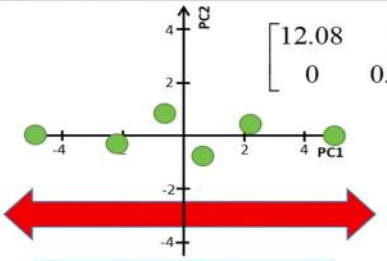
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

by the total variance,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

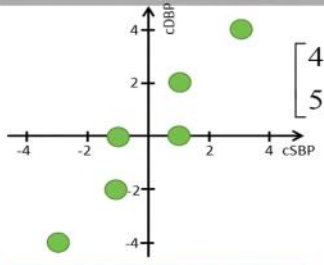
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\%\,\text{var} = \frac{12.08}{12.08+0.32} = 97.4\%$$

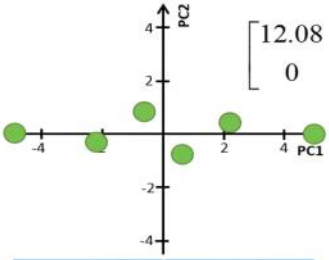
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

we see that the first principal component captures 97.4% of the total variance.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



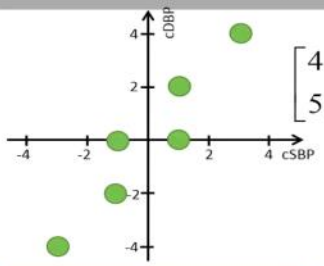
$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

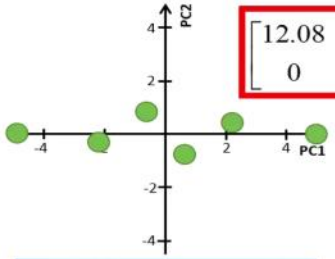
Note that the variances of the principal components correspond to the two eigenvalues we calculated earlier. Thus, the eigenvalues of the covariance matrix represent the variances of the principal components.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

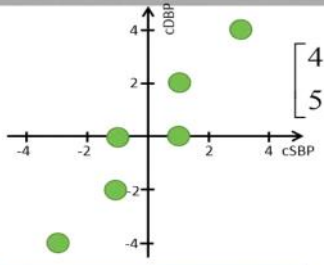


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

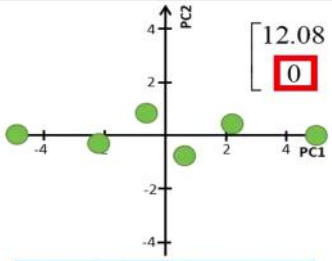
When we study the covariance matrix of our transformed data,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

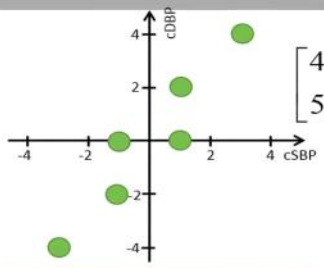


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

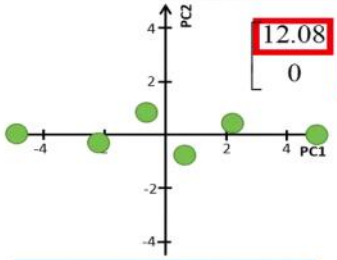
we see that the covariance between PC1 and PC2 is equal to zero, which means that PC1 and PC2 are completely uncorrelated.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

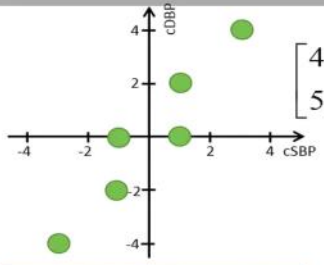


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

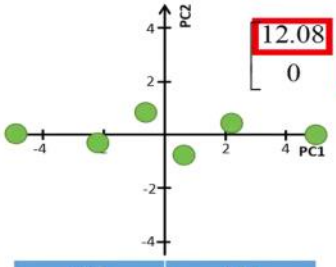
We also see that the variance of PC1 and PC2, correspond to the eigenvalues associated to the first and the second eigenvector.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

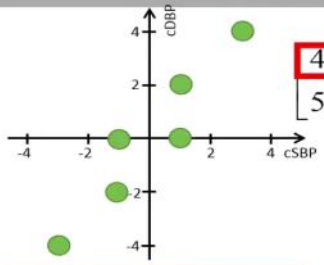


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

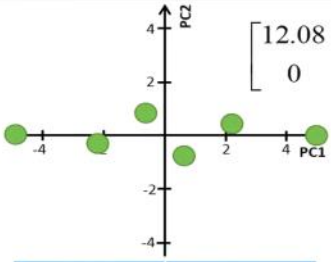
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Note that the total variance of PC1 and PC2 is about 12.4,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$



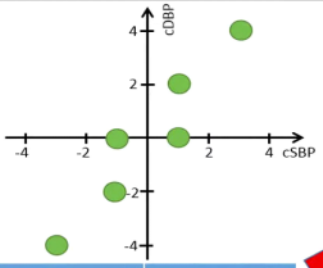
$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

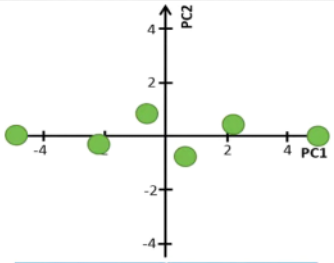
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

which corresponds to the total variance of the original variables.

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

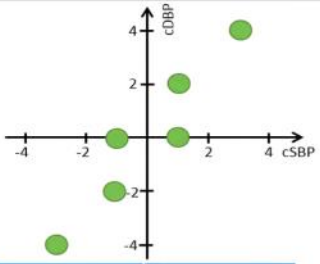


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

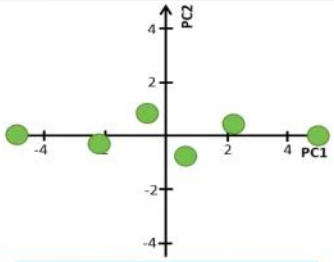
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Remember that when we multiplied our centered data,

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

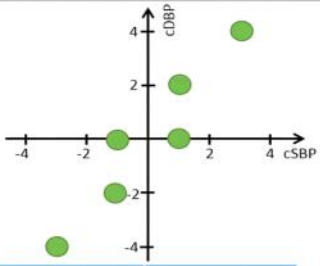


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

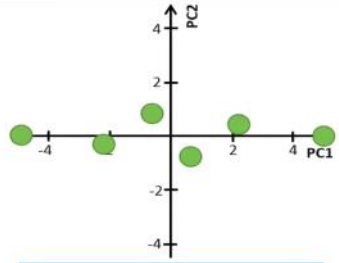
we got the transformed data.

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

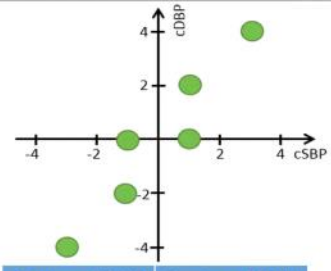


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

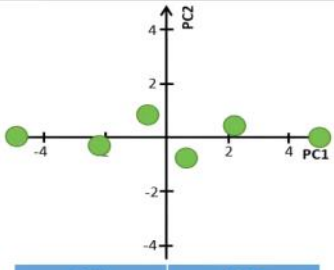
This is the same as using the following equation to calculate the principal components that we saw in the previous lecture,

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59\ cSBP + 0.81\ cDBP$$

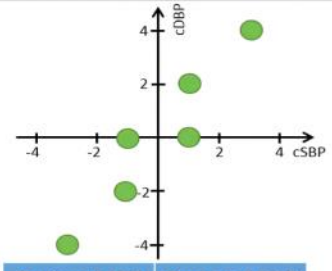


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

where the weights for the first principal component comes from the first eigenvector with the highest eigenvalue,

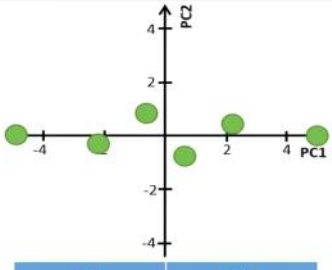
Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59\ cDBP$$

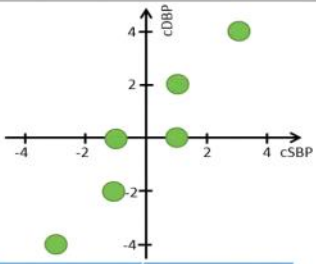


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

whereas the weights for the second principal component comes from the second eigenvector.

Interpret the PCA



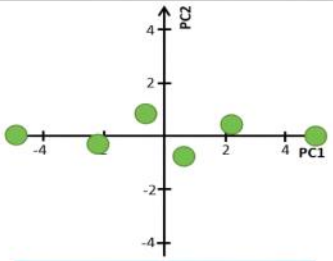
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

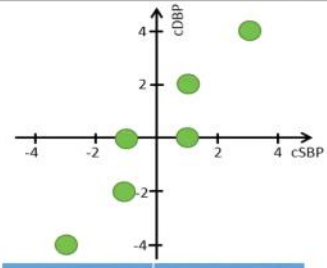
$$PC1_6 = 0.59 \cdot 3 + 0.81 \cdot 4 = 5$$



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

For example, if we would calculate the corresponding score for person number six,

Interpret the PCA



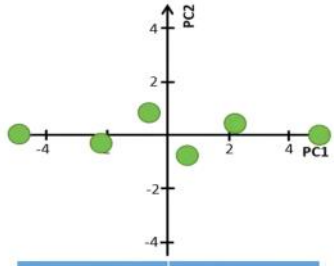
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

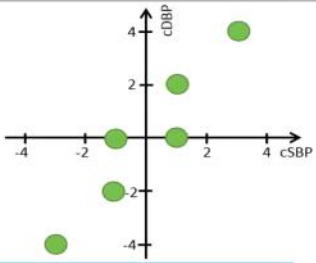
$$PC1_6 = 0.59 \cdot 3 + 0.81 \cdot 4 = 5$$



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

we would multiply the centered blood pressure values of person number six,

Interpret the PCA



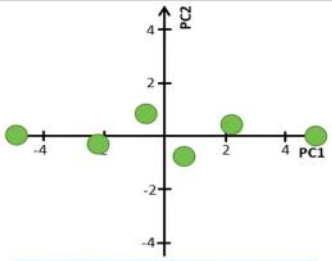
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

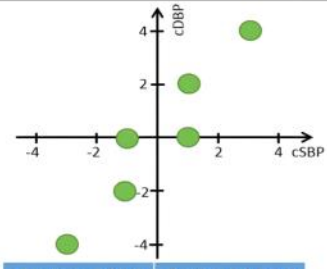
$$PC1_6 = 0.59 \cdot 3 + 0.81 \cdot 4 = 5$$



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

by the corresponding weights.

Interpret the PCA



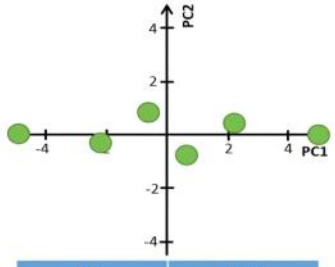
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

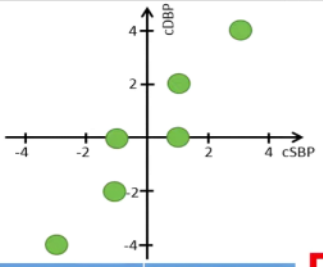
$$PC1_6 = 0.59 \cdot 3 + 0.81 \cdot 4 = 5$$



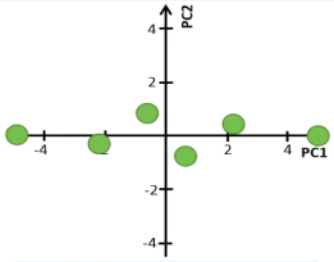
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

By adding these products, we would get a principal component score of 5.

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



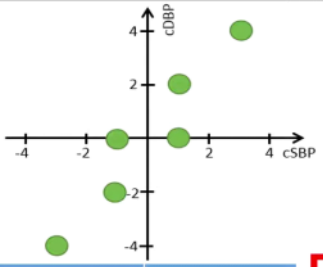
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

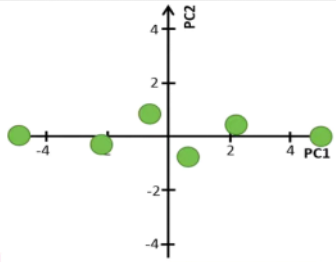
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Note that we can also use the following equation to calculate, for example, the first principal component,

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



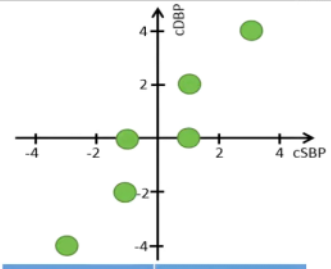
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

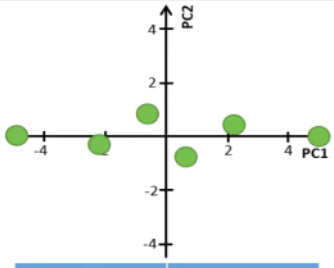
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Note that we can also use the following equation to calculate, for example, the first principal component,

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



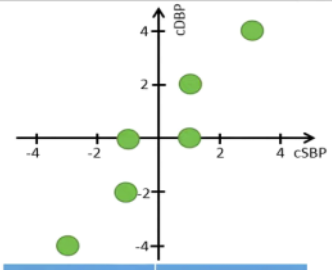
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

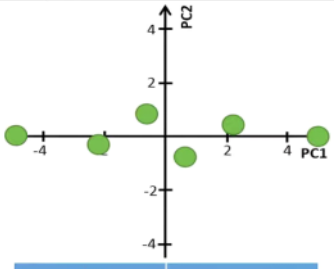
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

where we replace the variable for the centered data, by the variable of the original data minus its corresponding mean.

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



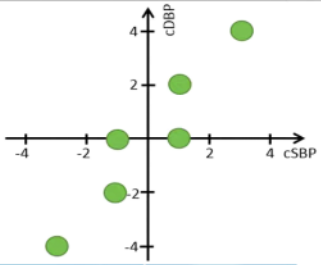
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

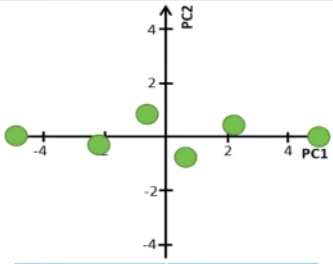
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

where we replace the variable for the centered data, by the variable of the original data minus its corresponding mean.

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

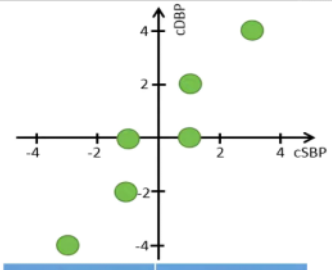
$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

$$PC1 = 0.59 \cdot (132 - 129) + 0.81 \cdot (86 - 82) = 5.0$$

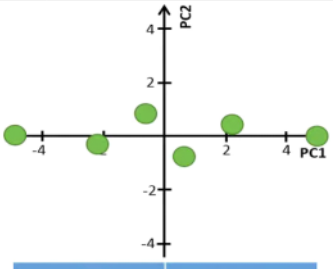
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

For example, if we would use the original blood pressure values for person number six,

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

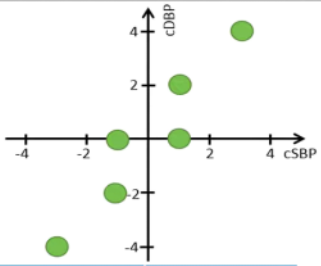
$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

$$PC1 = 0.59 \cdot (132 - 129) + 0.81 \cdot (86 - 82) = 5.0$$

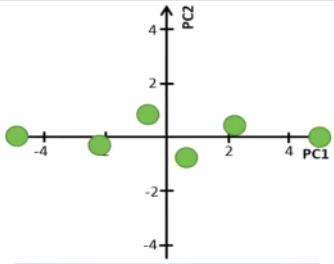
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

we would subtract the mean from the corresponding systolic and diastolic blood pressures,

Interpret the PCA



$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

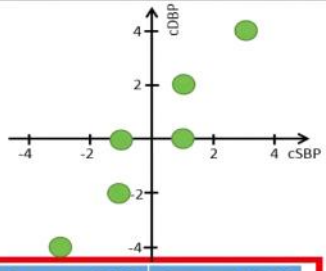
$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

$$PC1 = 0.59 \cdot (132 - 129) + 0.81 \cdot (86 - 82) = 5.0$$

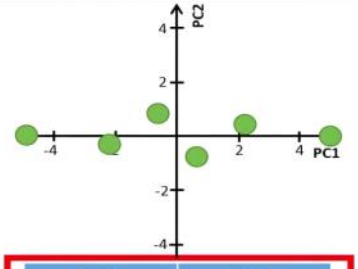
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

which would give the same values as for the centered data.

Interpret the PCA



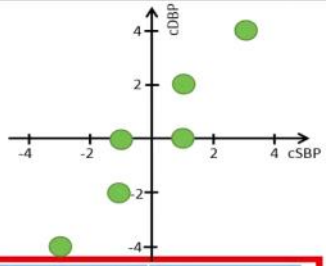
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0



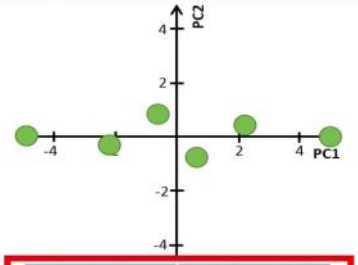
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Note that the general aim of using PCA is to reduce the dimensionality in the data. In other words, we like to reduce the number of variables we have.

Interpret the PCA



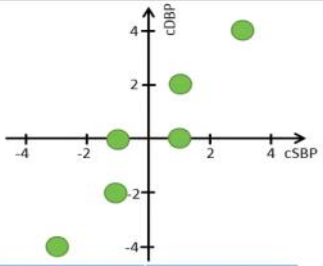
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

However, so far, we have not reduced the number of variables since we have the same number of principal components as the number of variables we started with.

Interpret the PCA

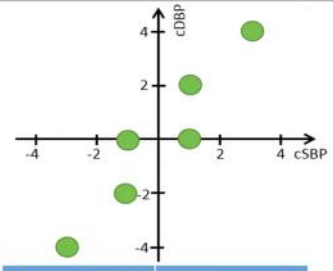


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Since the first principal component captures almost all variance, which can be interpreted as it stores almost all information about the two variables, we can simply delete the second principal component because it includes almost no information.

Interpret the PCA



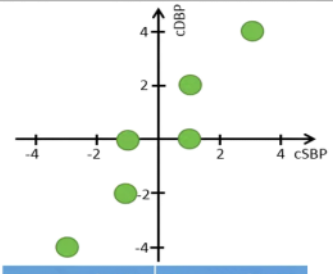
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

PC1
-5.0
-2.2
-0.6
0.6
2.2
5.0
Var=12.08

As we have seen previously, by using the following equation, we can combine the two variables into just one variable, in a way that maximize the variance of the linear combination.

Interpret the PCA



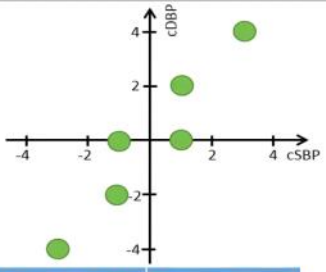
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

PC1
-5.0
-2.2
-0.6
0.6
2.2
5.0
Var=12.08

The weights can be interpreted as how much each variable contributes to the principal component.

Interpret the PCA



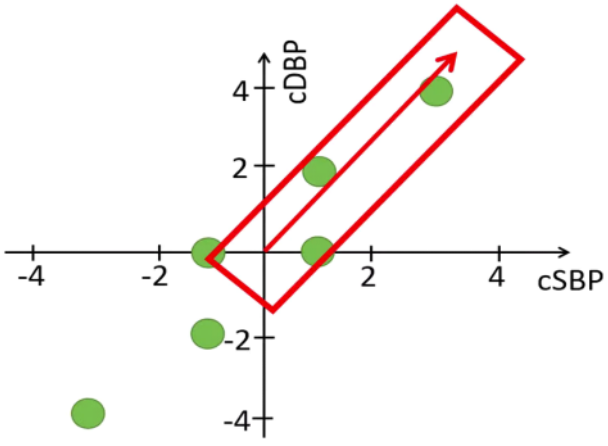
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot cSBP + \boxed{0.81}\ cDBP$$

PC1
-5.0
-2.2
-0.6
0.6
2.2
5.0
Var=12.08

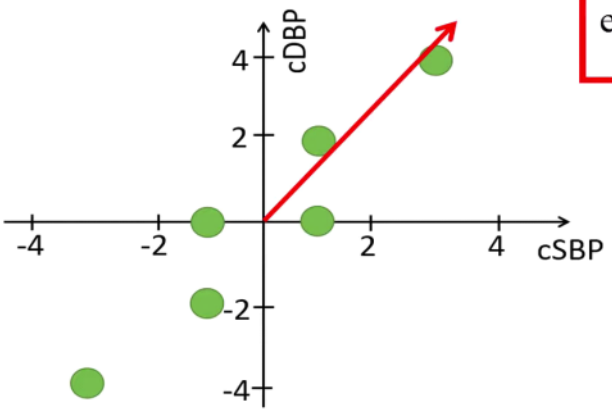
Since the absolute value of the weight for the diastolic blood pressure is higher than that of the systolic blood pressure, PCA put more weight on the diastolic blood pressure when the two variables are combined.

Interpret the eigenvector



Before we end this lecture, we will see why the first eigenvector points in the same direction as the data.

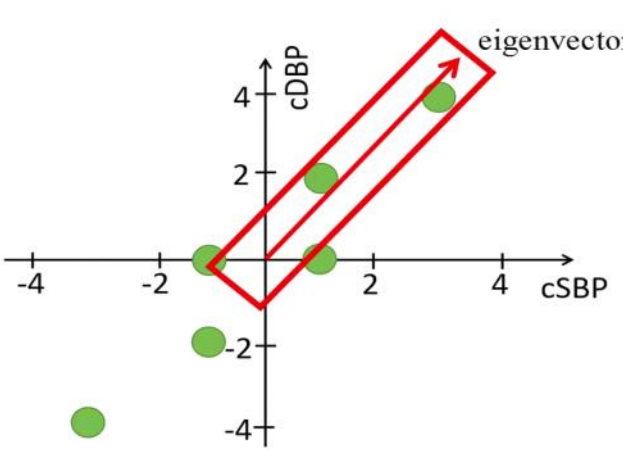
Interpret the eigenvector



$$\text{eigenvector} = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \cdot 6 = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

This is our previous eigenvector and if we extend it by multiplying by, for example, six,

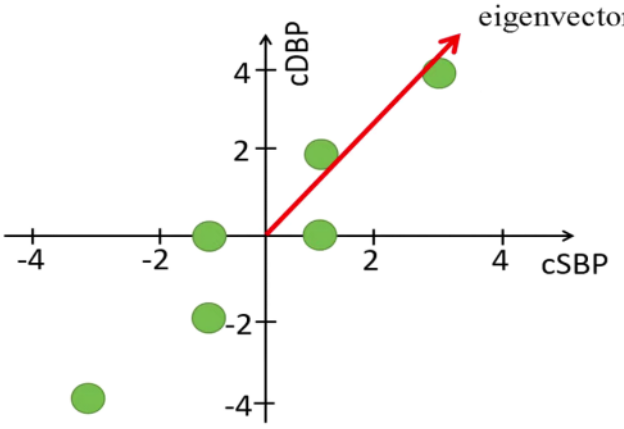
Interpret the eigenvector



$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

we can draw this vector. Note that this is also an eigenvector of the covariance matrix since it has the same direction as the eigenvector with the unit length.

Interpret the eigenvector

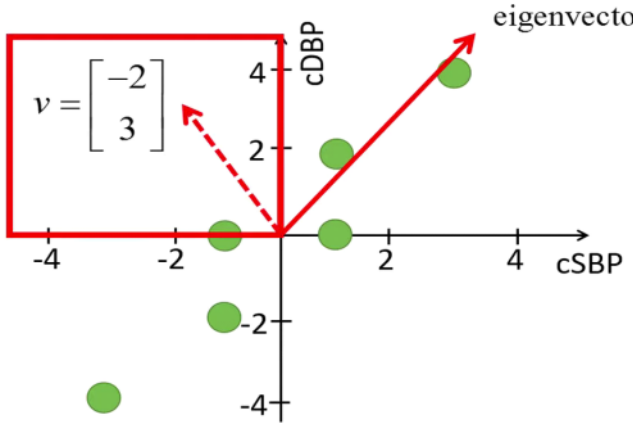


$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

This is our covariance matrix.

Interpret the eigenvector



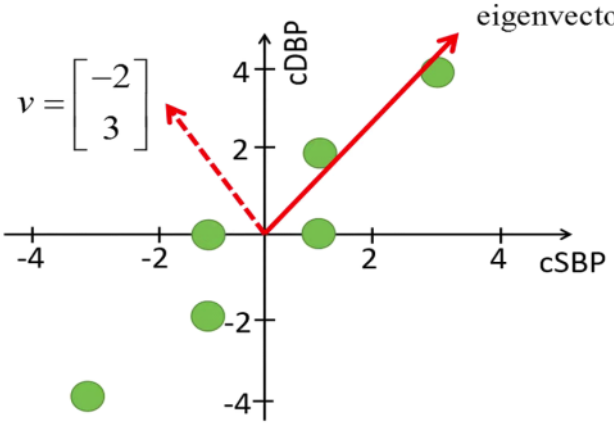
$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Let's take an arbitrary vector with the coordinates negative two and three.

Interpret the eigenvector



$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

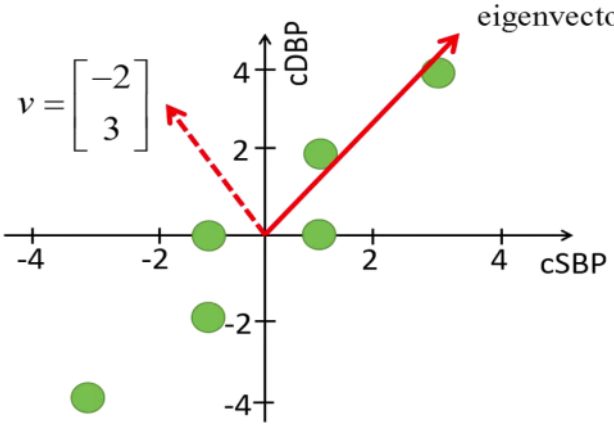
$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

If we would multiply the covariance matrix by this vector,

Interpret the eigenvector



$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

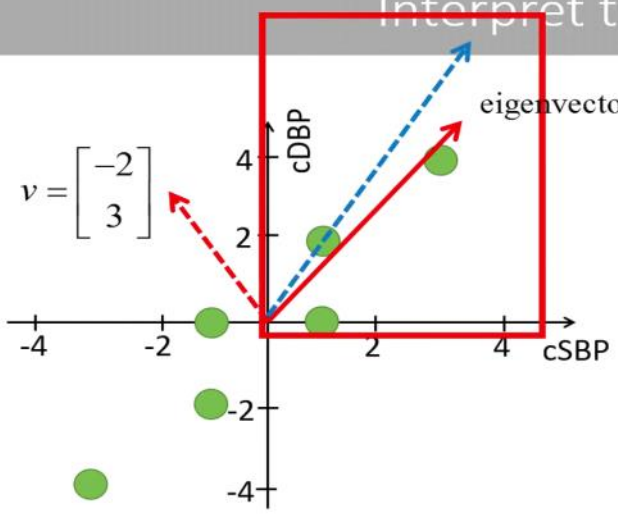
$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

we would get a new vector that has changed direction.

Interpret the eigenvector



$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

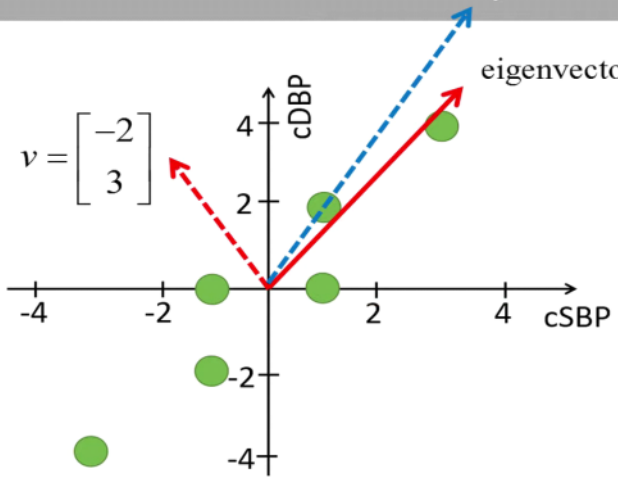
$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

We see that the covariance matrix transformed the vector so that it moved in a direction closer to the eigenvector. Note that we here do not plot the full length of the vector since it cannot fit the screen. The importance is its direction.

Interpret the eigenvector



$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

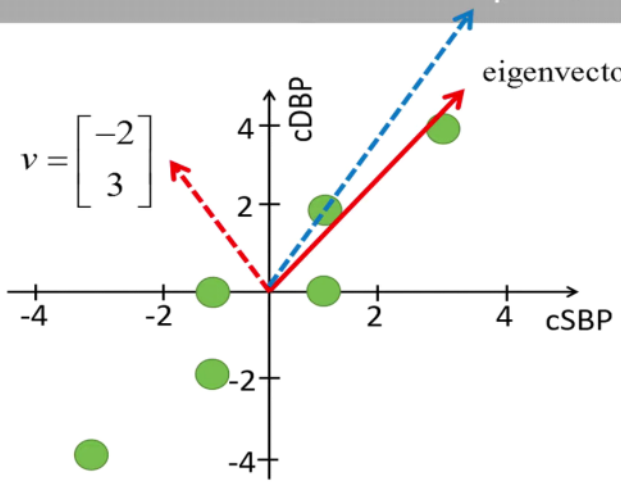
$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 12 \end{bmatrix} = \begin{bmatrix} 107 \\ 147 \end{bmatrix}$$

If we multiply the covariance matrix by this new vector,

Interpret the eigenvector



$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

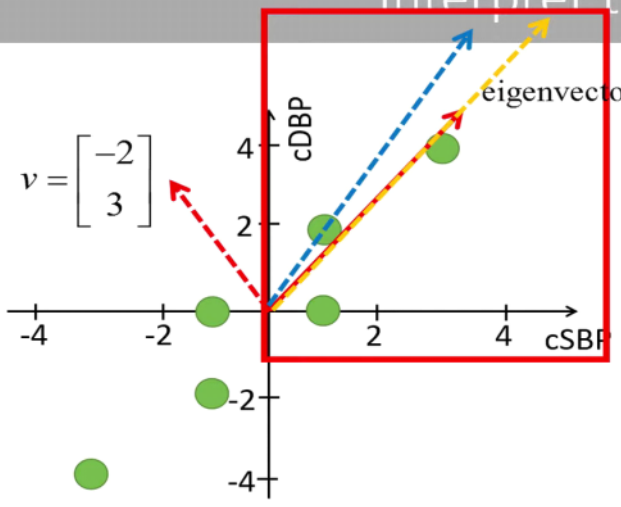
$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 12 \end{bmatrix} = \begin{bmatrix} 107 \\ 147 \end{bmatrix}$$

we will get a new vector again.

Interpret the eigenvector



$$v = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$\text{eigenvector} = \begin{bmatrix} 3.5 \\ 4.8 \end{bmatrix}$$

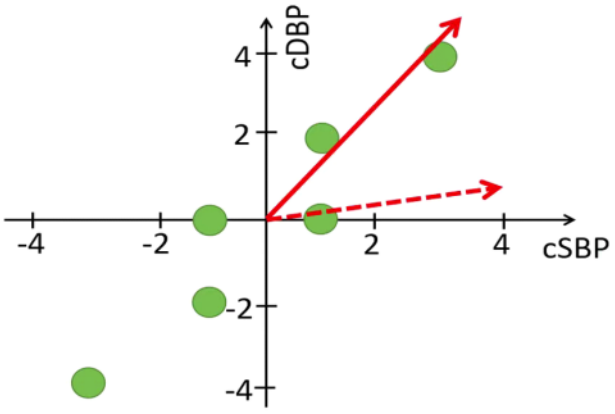
$$\text{cov}(cSBP, cDBP) = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 12 \end{bmatrix} = \begin{bmatrix} 107 \\ 147 \end{bmatrix}$$

This new vector will have more or less the same direction as the eigenvector.

Interpret the eigenvector

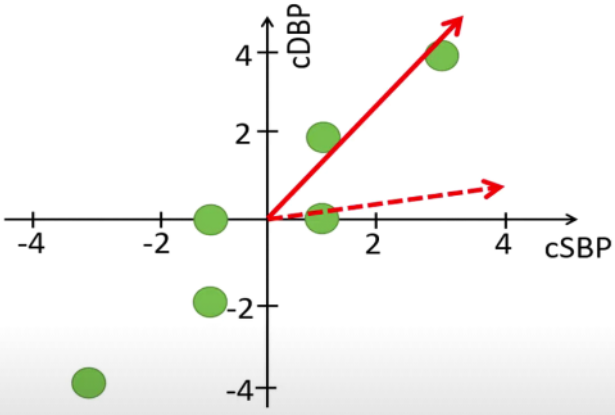


$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 23 \\ 30 \end{bmatrix}$$

Let's take another example vector, with the coordinate four and one.

PCA : the math - step-by-step with a simple example

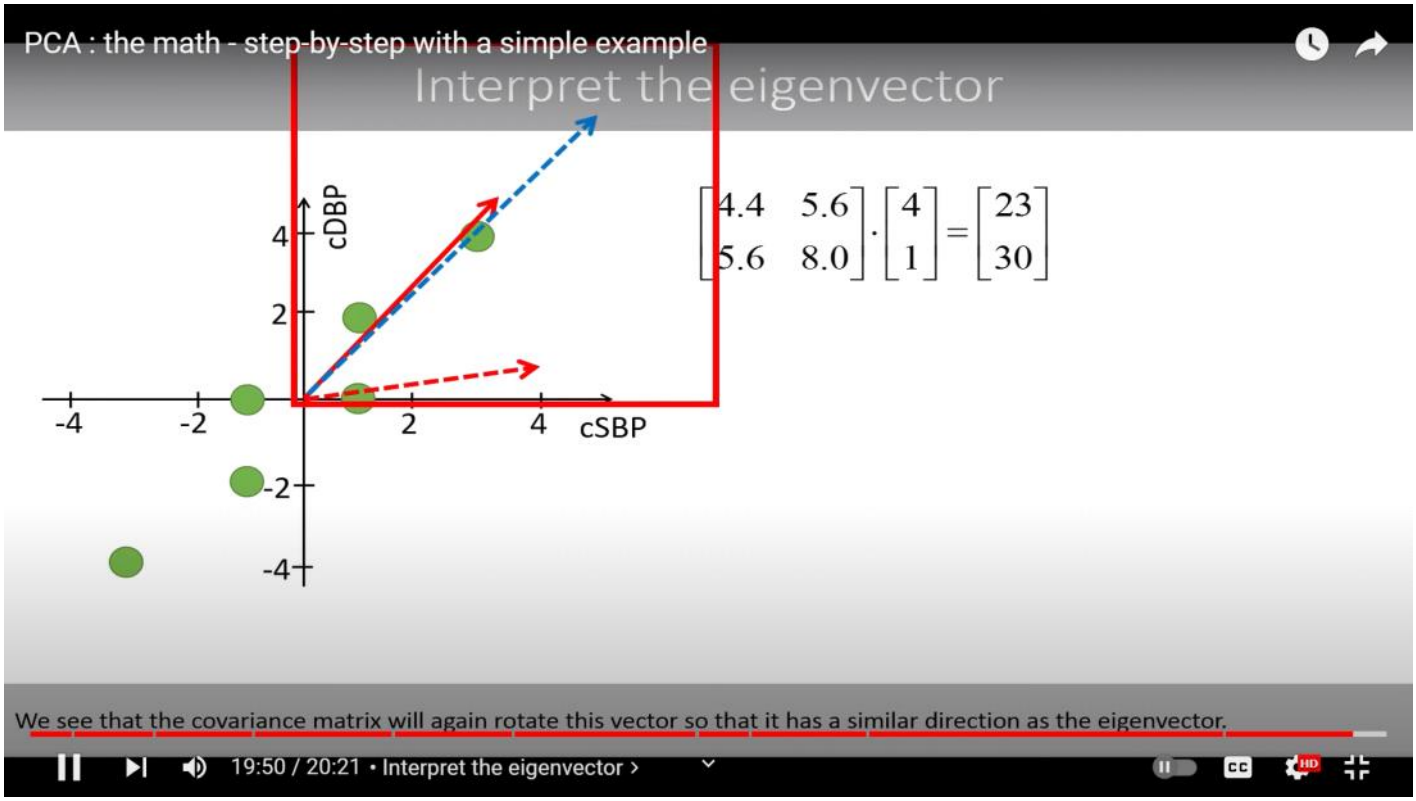
Interpret the eigenvector

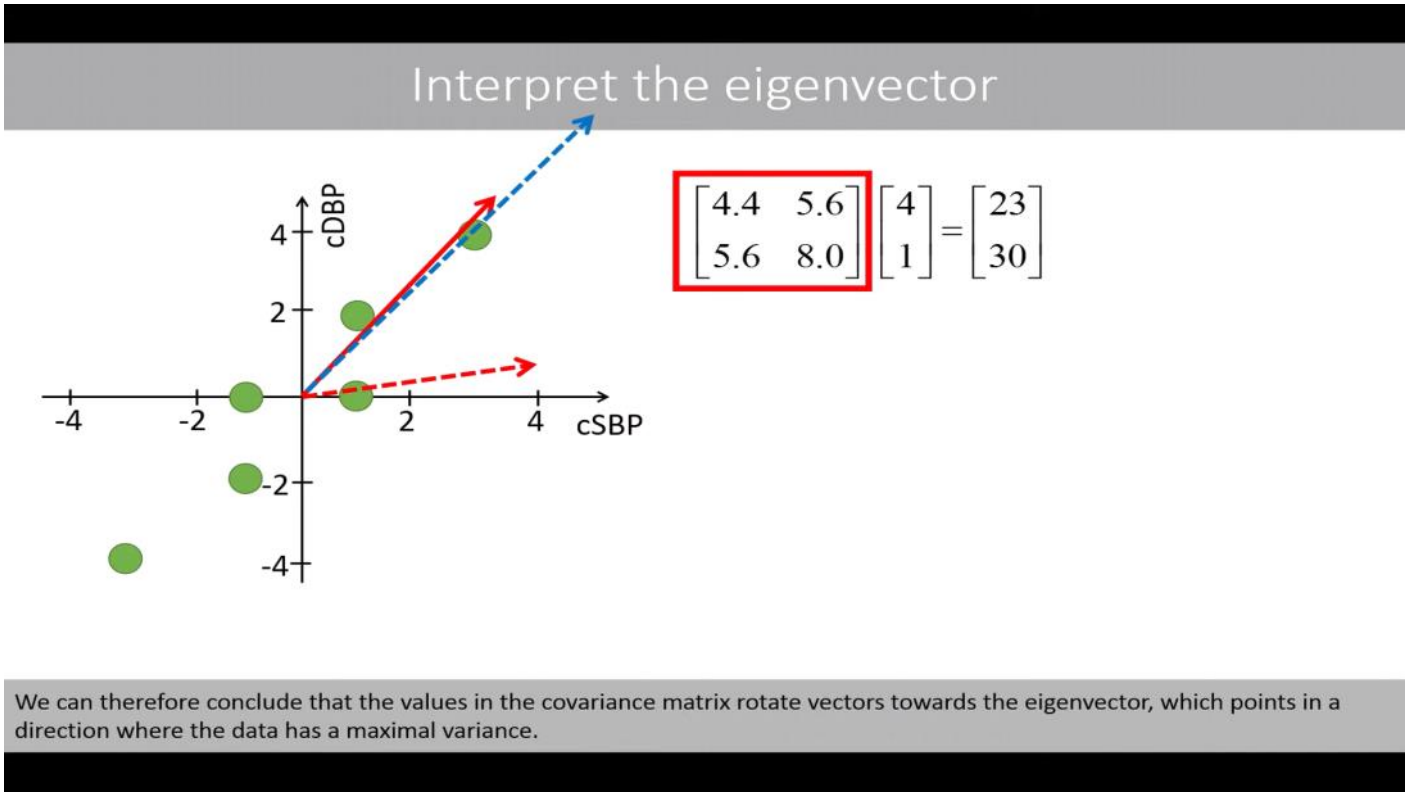


$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 23 \\ 30 \end{bmatrix}$$

Multiplying the covariance matrix by this arbitrary vector will result in the following vector.

19:33 / 20:21 • Interpret the eigenvector ›


PCA : the math - step-by-step with a simple example
Interpret the eigenvector
cDBP
cSBP
4
2
-2
-4
-4
-2
2
4
$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 23 \\ 30 \end{bmatrix}$
We see that the covariance matrix will again rotate this vector so that it has a similar direction as the eigenvector.
19:50 / 20:21 • Interpret the eigenvector ›


Interpret the eigenvector
cDBP
cSBP
4
2
-2
-4
-4
-2
2
4
$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 23 \\ 30 \end{bmatrix}$
We can therefore conclude that the values in the covariance matrix rotate vectors towards the eigenvector, which points in a direction where the data has a maximal variance.