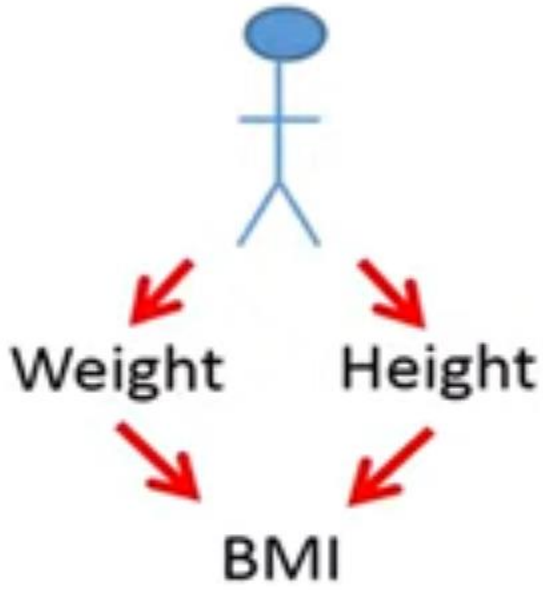
PCA Made Easy

Zahoor Tanoli (PhD)

CUI, Attock Campus

Combining variables

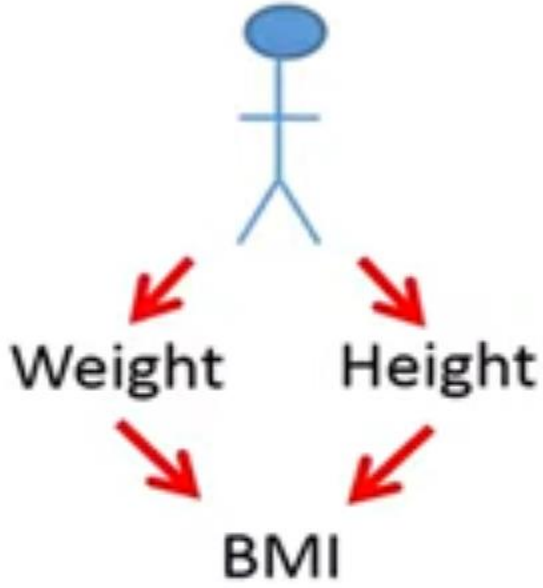


Cholesterol = Weight + Height

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

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



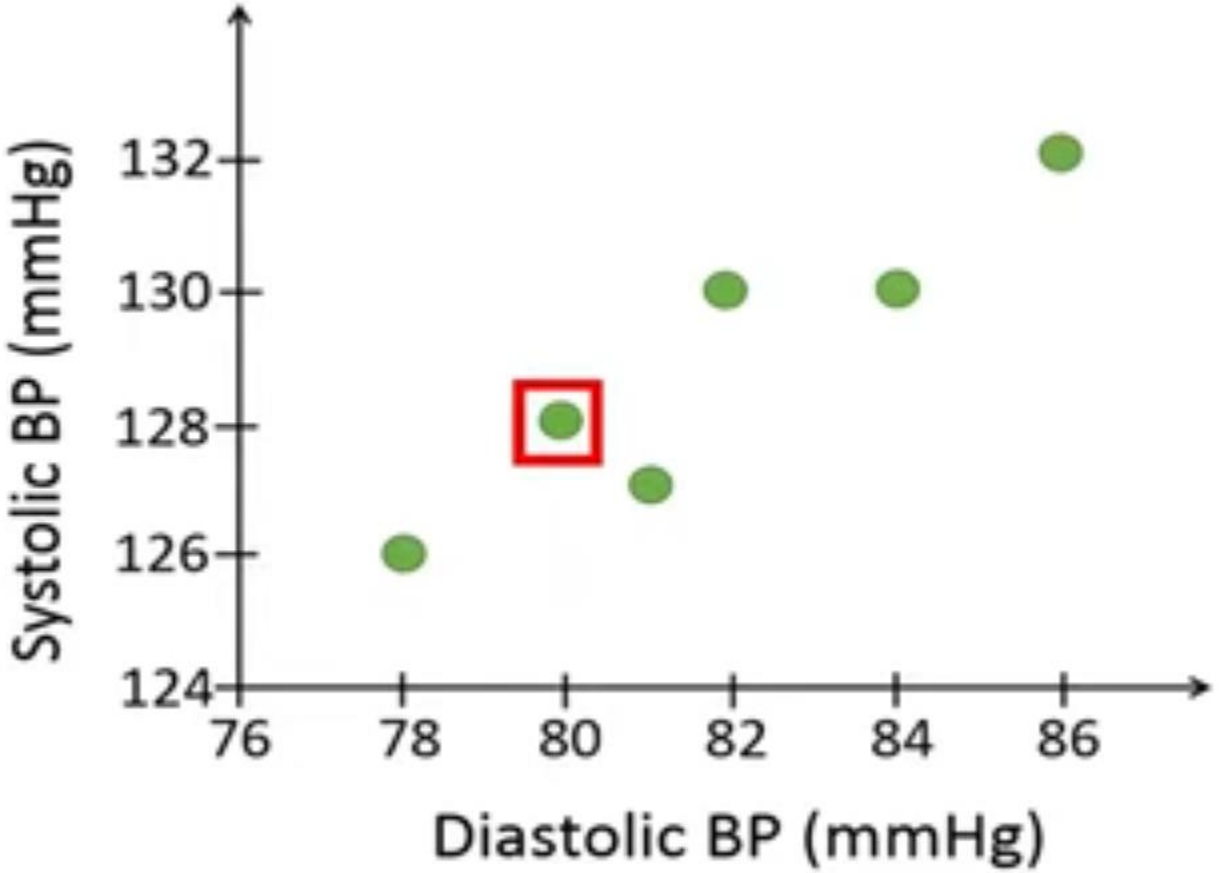
Cholesterol = Weight + Height

Cholesterol = BMI

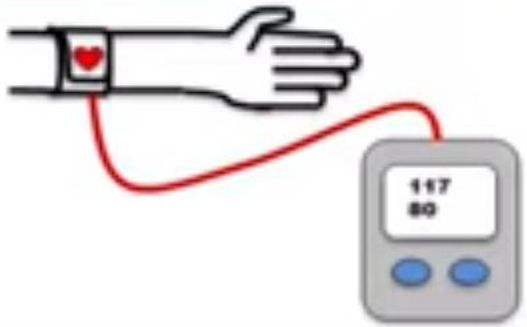
$$BMI = \frac{Weight_{kg}}{Height_{m}^{2}}$$

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

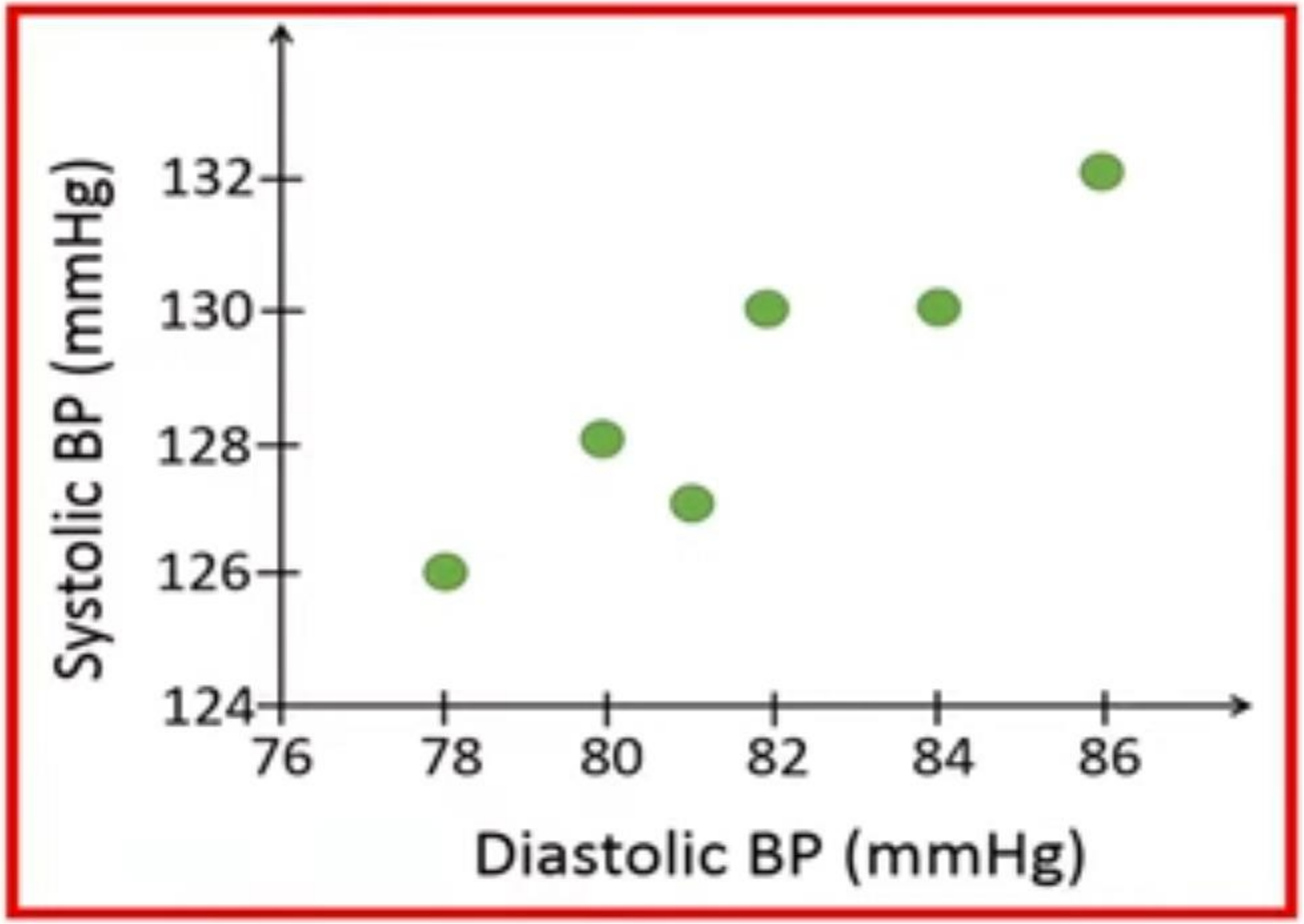


Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

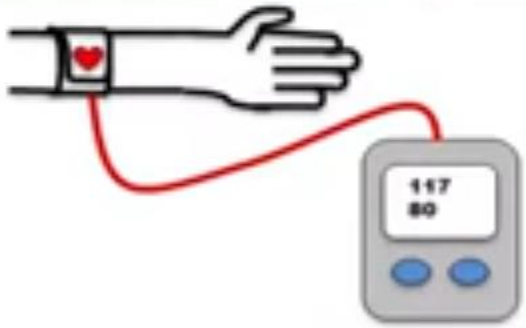


Person number two has a diastolic blood pressure of 80 and a systolic blood pressure of 128, and so on.

Combining variables

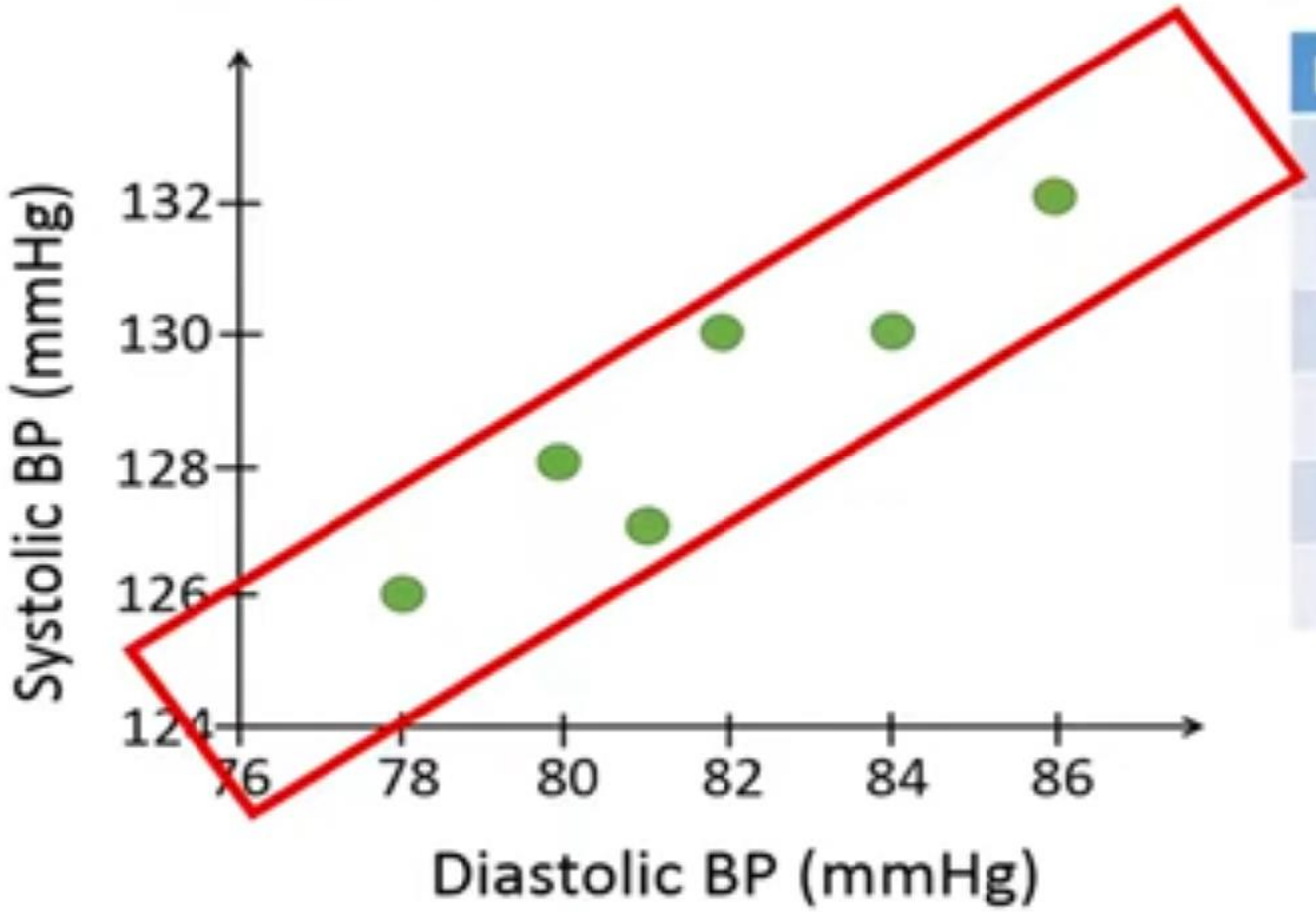


Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132

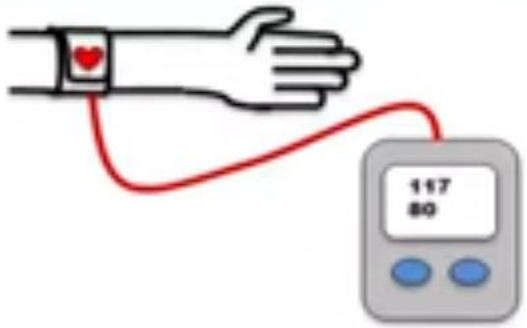


For this data set, there seems to be a strong positive correlation between the upper and lower blood pressure. If a person has a high systolic blood pressure, it is likely that the person also has a high diastolic blood pressure.

Combining variables



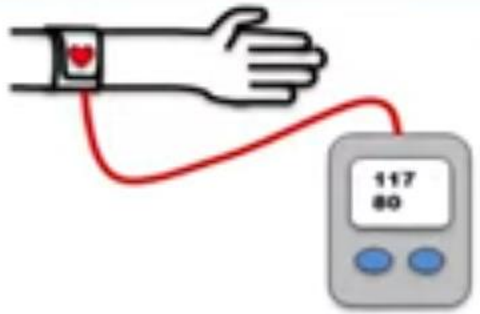
Diastolic BP	Systolic BP
78	126
80	128
81	127
82	130
84	130
86	132



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

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

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

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

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

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

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

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	104
82	130	106
84	130	107
86	132	109

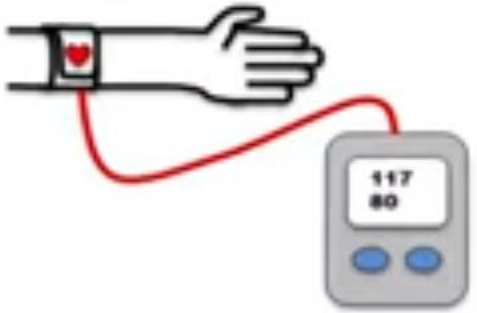
$$BP = \alpha_1 DBP + \alpha_2 SBP$$

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	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

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

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.

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

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

PCA

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

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

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

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

$$BP = 0.8DBP + 0.6SBP$$

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**

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

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

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.

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

	DBP	SBP
DBP	8.17	5.97
SBP	5.97	4.97

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

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

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

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

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

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	...
...	..	..	..	..	..	..	...

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	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
...	..	..

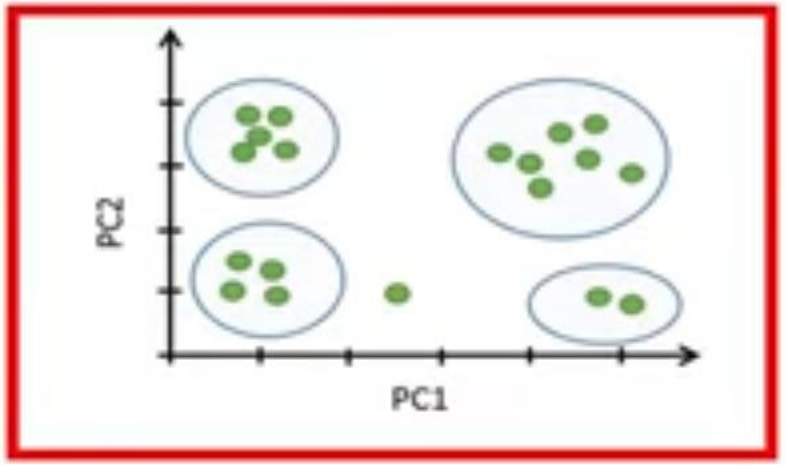
When we combine variables with PCA, we will get these kinds of scores that are centered around zero, which explains why about half of the values are negative.

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,

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

PCA

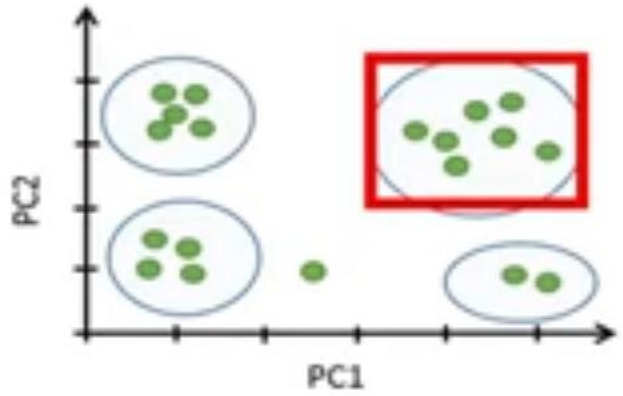
Person	PC1	PC2
1	3.4	4.5
2	3.7	4.4
3	-21.2	-15.2
4	-20.2	-16.5
5	-8.4	-8.9
6	-0.2	10.2
...	..	..



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
...	..	..



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

What is PCA?

- **Principal component analysis** (**PCA**) is a statistical procedure that is used to reduce the dimensionality
- Converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables
- Converted uncorrelated variables are called principal components
- It is often used as a dimensionality reduction technique.

PCA Steps

1. Standardize the dataset
2. Calculate the covariance matrix for the features in the dataset
3. Calculate the eigenvalues and eigenvectors for the covariance matrix
4. Sort eigenvalues and their corresponding eigenvectors
5. Pick k eigenvalues and form a matrix of eigenvectors
6. Transform the original matrix

Standardize the Dataset

- We have the dataset which has 4 features and a total of 5 training examples

f1	f2	f3	f4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

Step-1 (Standardization)

- Calculate the mean and standard deviation for each feature

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

	f1	f2	f3	f4
μ =	4	3	3	3.4
σ =	3	1.58114	1.73205	2.30217

$$x_{new} = \frac{x - \mu}{\sigma}$$

f1	f2	f3	f4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

Step-1 (Standardization)

- Each feature in the dataset is transformed using $x_{new} = \frac{x - \mu}{\sigma}$

	f1	f2	f3	f4
μ =	4	3	3	3.4
σ =	3	1.58114	1.73205	2.30217

f1	f2	f3	f4
-1	-0.63246	0	0.26062
0.33333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.33333	0	-0.57735	-1.04249
1.33333	-1.26491	-0.57735	-0.60812

Step-2: Covariance Matrix

- Formula to calculate the covariance matrix:

For Population

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

- Covariance matrix for the given dataset will be calculated as:

	f1	f2	f3	f4
f1	var(f1)	cov(f1,f2)	cov(f1,f3)	cov(f1,f4)
f2	cov(f2,f1)	var(f2)	cov(f2,f3)	cov(f2,f4)
f3	cov(f3,f1)	cov(f3,f2)	var(f3)	cov(f3,f4)
f4	cov(f4,f1)	cov(f4,f2)	cov(f4,f3)	var(f4)

Step-2: Covariance Matrix

- For standardized dataset, mean for each feature is 0 and the standard deviation is 1
- var(f1) = ((-1.0-0)² + (0.33-0)² + (-1.0-0)² +(0.33–0)² +(1.33–0)²)/5
- **var (f1) = 0.8**

- cov(f1,f2) = ((-1.0–0)*(-0.632456-0) +
  (0.33–0)*(1.264911-0) +
  (-1.0–0)* (0.632456-0)+
  (0.33–0)*(0.000000 -0)+
  (1.33–0)*(-1.264911–0))/5
- **cov(f1,f2 = -0.25298**

f1	f2	f3	f4
-1	-0.63246	0	0.26062
0.33333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.33333	0	-0.57735	-1.04249
1.33333	-1.26491	-0.57735	-0.60812

Final Covariance Matrix

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

Step-3: Calculate Eigenvalues and Eigenvectors

- Eigenvector is a nonzero vector that changes at most by a scalar factor
- eigenvalue is the factor by which the eigenvector is scaled
- Let A be a square matrix (in our case the covariance matrix)
  - v a vector and λ a scalar that satisfies $Av = \lambda v$
  - λ is called eigenvalue associated with eigenvector v of A
- Rearranging the equation as: $Av-\lambda v =0$; $(A-\lambda I)v = 0$
- Since v is a non-zero vector then only way to get this equation equal to zero is: **$\det(A-\lambda I) = 0$**

Eigenvalues

A-λI =

	f1	f2	f3	f4
f1	0.8 - λ	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8- λ	0.51121	0.4945
f3	0.03849	0.51121	0.8 - λ	0.75236
f4	-0.14479	0.4945	0.75236	0.8 - λ

- Solving for **det(A-λI) = 0**
- **λ** = 2.51579324 , 1.0652885 , 0.39388704 , 0.02503121
- **Note: Values are in the sorted order so sorting step will be ignored**

Eigenvectors

- Solving the (A-λI)v = 0 equation for v vector with different λ values

$$\begin{pmatrix} 0.800000-\lambda & -(0.252982) & 0.038490 & -(0.144791) \\ -(0.252982) & 0.800000-\lambda & 0.511208 & 0.494498 \\ 0.038490 & 0.511208 & 0.800000-\lambda & 0.752355 \\ -(0.144791) & 0.494498 & 0.752355 & 0.800000-\lambda \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = 0$$

- For λ = 2.51579324, solving using Cramer's rule, the values for v vector are
- v1 = 0.16195986
- v2 = -0.52404813
- v3 = -0.58589647
- v4 = -0.59654663

Eigenvectors

- Using the values λ = 2.51579324 , 1.0652885 , 0.39388704 , and 0.02503121

$$\begin{pmatrix} 0.800000-\lambda & -(0.252982) & 0.038490 & -(0.144791) \\ -(0.252982) & 0.800000-\lambda & 0.511208 & 0.494498 \\ 0.038490 & 0.511208 & 0.800000-\lambda & 0.752355 \\ -(0.144791) & 0.494498 & 0.752355 & 0.800000-\lambda \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = 0$$

e1	e2	e3	e4
0.161960	-0.917059	-0.307071	0.196162
-0.524048	0.206922	-0.817319	0.120610
-0.585896	-0.320539	0.188250	-0.720099
-0.596547	-0.115935	0.449733	0.654547

Step-5: Pick k top eigenvalues

- If choosing the top 2 eigenvectors, the matrix will be:

e1	e2	e3	e4
0.161960	-0.917059	-0.307071	0.196162
-0.524048	0.206922	-0.817319	0.120610
-0.585896	-0.320539	0.188250	-0.720099
-0.596547	-0.115935	0.449733	0.654547

e1	e2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935

Transform the original matrix

- Feature matrix * top k eigenvectors = Transformed Data

f1	f2	f3	f4
-1.000000	-0.632456	0.000000	0.260623
0.333333	1.264911	1.732051	1.563740
-1.000000	0.632456	-0.577350	-0.173749
0.333333	0.000000	-0.577350	-1.042493
1.333333	-1.264911	-0.577350	-0.608121

(5,4)

*

e1	e2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935

(4,2)

=

nf1	nf2
0.014003	0.755975
-2.556534	-0.780432
-0.051480	1.253135
1.014150	0.000239
1.579861	-1.228917

(5,2)