

Topic 2: A Look at Multivariate Techniques

Dr. Lee Fawcett

Thursday, September 23rd, 2021

Introduction

Introduction

Multivariate Data Analysis, often abbreviated to MVA, provides the foundations of some of the buzz phrases being used in modern data applications:

- Data Science
- Big Data Analytics
- Machine Learning
- Artificial Intelligence

Over the course of the next 50 minutes we will have a whistle-stop tour of MVA, focusing primarily on the 'workhorse' of modern MVA techniques: **Principal Components Analysis**.

Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – often **two steps**:

- Exploratory Data Analysis (building upon what we've just seen with Joe in the first session)
- Modelling (what we'll cover in one of tomorrow's sessions)

Why 'multivariate'?

If you could solve all problems by taking a single measurement on a system, the world would be a much simpler place!

However, complex systems often require multiple measurements, on many variables, to understand them.

Example: **Predicting the weather**. Meteorologists require information on pressure, wind conditions, humidity, season, ...

Why 'multivariate'?

Multivariate Data Analysis, often abbreviated to MVA, is the investigation of many variables, simultaneously, in order to understand the relationships that might exist between them.

An MVA might be as simple as looking at the relationship between two variables through a scatterplot, right up to analysing **millions of variables together**.

Why 'multivariate'?

An MVA can be **highly graphical**, although the usual output graphs might not be those you are used to seeing.

For example, you might be interested in the association between height and weight for a cohort of students; this could easily be explored through a 2-D scatterplot.

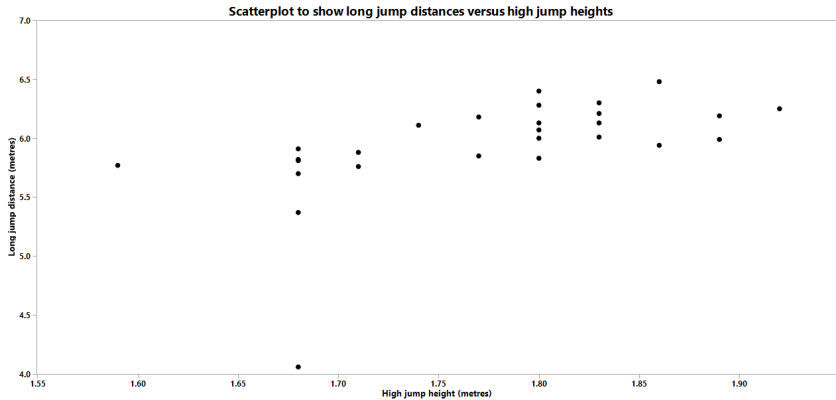
Throw a third variable into the mix – perhaps shoe size – and we're into the realms of 3-D scatterplots.

Graphically displaying the results for more than three variables **requires some ingenuity**, since we live in a 3-D world!

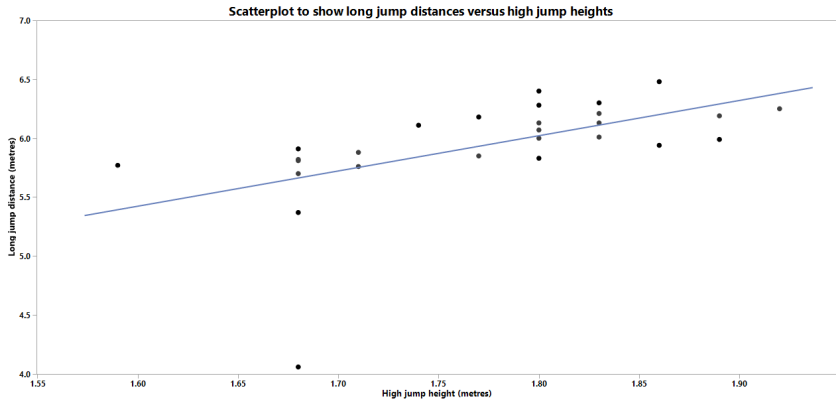
Example: A look at the heptathlon dataset

	Name	100M hurdles	High jump	Shot put	200M	Long jump	Javelin	800M
	Jessica Ennis	12.54	1.86	14.28	22.83	6.48	47.49	128.65
	Lilli Schwarzkopf	13.26	1.83	14.77	24.77	6.30	51.73	130.50
	Austra Skujyte	14.00	1.92	17.31	25.43	6.25	51.13	140.59
	Antoinette Nana Djimou Ida	12.96	1.80	14.26	24.72	6.13	55.87	133.62
	Jessica Zelinka	12.65	1.68	14.81	23.32	5.91	45.75	129.15
	Kristina Savitskaya	13.37	1.83	14.77	24.46	6.21	43.70	132.27
	Laura Ikauniece	13.71	1.83	12.64	24.16	6.13	51.27	132.13
	Hanna Melnychenko	13.32	1.80	12.96	24.09	6.40	43.86	132.90
	Brianne Theisen	13.30	1.83	12.89	24.35	6.01	46.47	129.27
	Dafne Schippers	13.48	1.80	13.67	22.83	6.28	36.63	135.52
	Nadine Broersen	13.64	1.86	13.57	25.13	5.94	51.98	136.98
	Jessica Samuelsson	13.58	1.77	14.18	24.25	6.18	42.02	131.31
	Katarina Johnson-Thompson	13.48	1.89	11.32	23.73	6.19	38.37	130.76
	Sharon Day	13.57	1.77	14.28	24.36	5.85	43.90	131.31
	Yana Maksimava	13.97	1.89	14.09	25.43	5.99	42.33	133.37
	Eliska Klucinova	14.01	1.80	12.93	25.00	6.13	45.65	136.08
	Ellen Sprunger	13.35	1.71	12.62	23.59	5.88	45.63	137.54
	Olga Kurban	13.46	1.80	13.71	23.88	5.83	40.36	139.82
	Marisa De Aniceto	13.74	1.71	13.09	25.26	5.76	51.98	136.20
	Gyorgyi Farkas	14.33	1.80	13.55	25.72	6.07	46.52	137.83
	Grit Sadeiko	13.50	1.74	12.43	24.25	6.11	44.12	143.01
	Sofia Ifadidou	13.82	1.68	12.96	25.91	5.81	56.96	142.03
	Ivona Dacic	14.58	1.80	12.19	24.29	6.00	41.82	135.90
	Sarah Cowley	13.95	1.80	12.37	25.60	6.00	41.90	139.01
	Louise Hazel	13.48	1.59	12.81	24.48	5.77	47.38	138.78
	Ida Marcussen	14.08	1.68	14.26	25.15	5.82	42.26	133.62
	Chantae McMillan	13.49	1.68	14.92	25.25	5.37	49.78	160.55
	Julia Machtig	14.54	1.68	14.99	25.38	4.06	44.40	140.34
	Irina Karpova	14.21	1.68	11.68	25.42	5.70	35.75	148.93

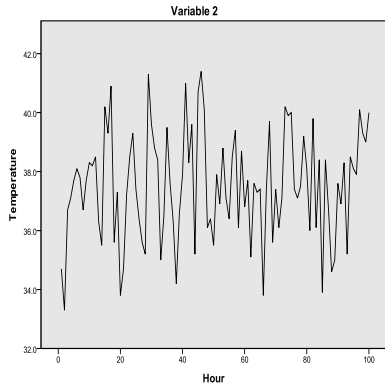
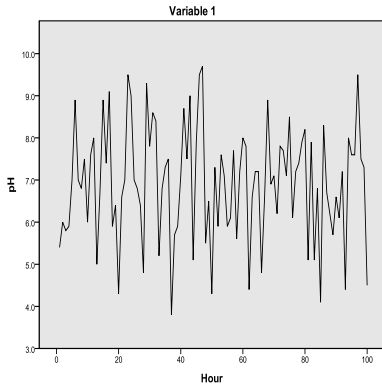
Example: A look at the heptathlon dataset



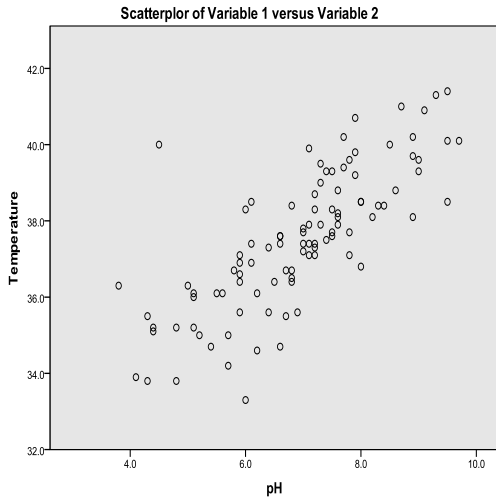
Example: A look at the heptathlon dataset



Example: A hidden outlier!



Example: A hidden outlier!



How to proceed?

The first step of any multivariate analysis should be **exploration**:

- Are there any dependencies between our variables?
- Are there any outliers, or groupings in our data, that cannot be seen by looking at each of the variables individually?

If there *are* patterns in our data, we might then want to **exploit** these patterns.

That will lead us on to what is often the second step of any multivariate analysis: **modelling**.

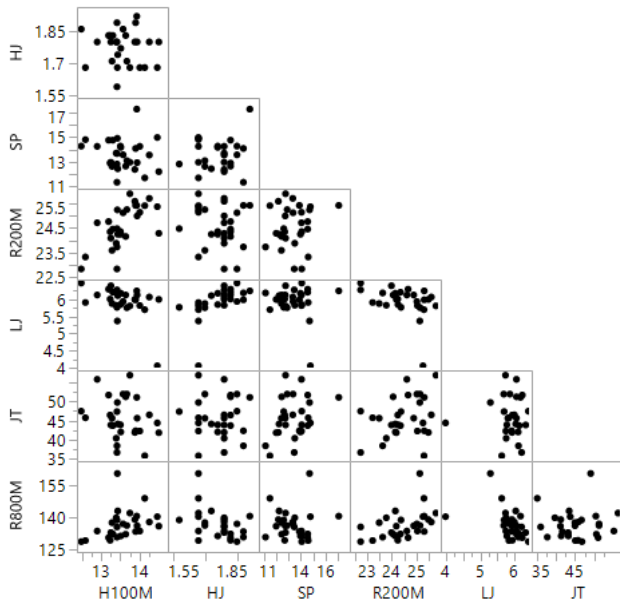
Simple exploration

Simple methods for summarising multivariate datasets

Before embarking on any rigorous analysis of a multivariate dataset (or *any* dataset, for that matter!), **plot it!**

- Univariate datasets: histograms, bar charts/pie charts, ...
- Bivariate datasets: histograms for the 'margins', **scatterplots** for the dependence
- Multivariate datasets...?

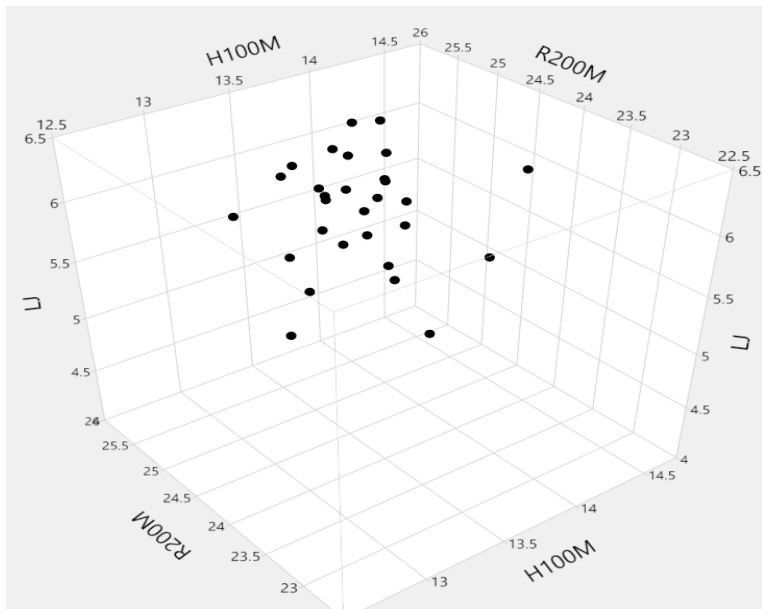
Scatterplot matrices



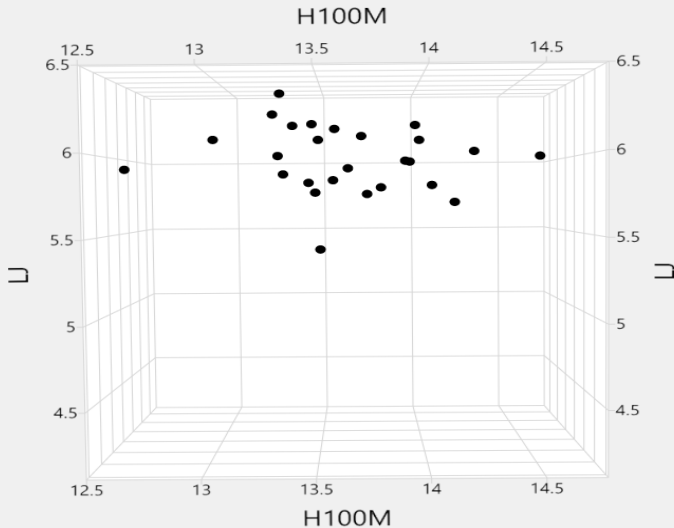
Correlation matrices

Multivariate							
Correlations							
	H100M	HJ	SP	R200M	LJ	JT	R800M
H100M	1.0000	-0.0842	-0.1419	0.6567	-0.4435	-0.2221	0.3590
HJ	-0.0842	1.0000	0.1575	-0.1353	0.5498	-0.0181	-0.4277
SP	-0.1419	0.1575	1.0000	0.1083	-0.1172	0.3268	-0.0023
R200M	0.6567	-0.1353	0.1083	1.0000	-0.3779	0.3250	0.4607
LJ	-0.4435	0.5498	-0.1172	-0.3779	1.0000	0.0102	-0.4485
JT	-0.2221	-0.0181	0.3268	0.3250	0.0102	1.0000	0.0306
R800M	0.3590	-0.4277	-0.0023	0.4607	-0.4485	0.0306	1.0000

3-D Scatterplots



3-D Scatterplots



Other methods

There are a variety of other methods available for summarising multivariate datasets both graphically and numerically. These include:

- **Star symbol plots:** Each individual is represented by a star, these stars having the same number of rays as we have variables. The length of the rays are proportional to the measurements made on each variable
- **Mahalanobis distance:** Put simply, this is the distance between two points in multivariate space. It is a numerical summary often used to detect outliers in multivariate space

Principal Components Analysis

Principal Components Analysis

- **Principal Components Analysis** (PCA) is the analysis of variability in a multivariate dataset
- PCA is often referred to as the “**workhorse of multivariate methods**”
- PCA provides some powerful graphical tools for exploratory analysis, and enables datasets with many variables to be displayed graphically in 2-D space
- Thus, PCA is referred to as a **dimension reduction** technique

Principal Components Analysis

- The main aim is to find new variables which are linear combinations of the original variables
- The new variables are such that a few of them will explain most of the variability in the data, hence effectively **reducing the number of variables**
- Unlike **stepwise regression**, PCA includes information on *all* variables through these new variables

PCA of the heptathlon dataset

Recall we have seven variables, which we will now label:

X_1 = Time in 100 metre hurdles

X_2 = Height in high jump

X_3 = Distance in shot put

X_4 = Time in 200 metres

X_5 = Distance in long jump

X_6 = Distance thrown in javelin

X_7 = Time in 800 metres

PCA of the heptathlon dataset

The aim of a PCA is to **construct new variables**:

$$Y_1 = a_{1,1}X_1 + a_{2,1}X_2 + a_{3,1}X_3 + a_{4,1}X_4 + a_{5,1}X_5 + a_{6,1}X_6 + a_{7,1}X_7$$

$$Y_2 = a_{1,2}X_1 + a_{2,2}X_2 + a_{3,2}X_3 + a_{4,2}X_4 + a_{5,2}X_5 + a_{6,2}X_6 + a_{7,2}X_7$$

$$Y_3 = a_{1,3}X_1 + a_{2,3}X_2 + a_{3,3}X_3 + a_{4,3}X_4 + a_{5,3}X_5 + a_{6,3}X_6 + a_{7,3}X_7$$

$$Y_4 = a_{1,4}X_1 + a_{2,4}X_2 + a_{3,4}X_3 + a_{4,4}X_4 + a_{5,4}X_5 + a_{6,4}X_6 + a_{7,4}X_7$$

$$\vdots \quad \quad \vdots$$

PCA of the heptathlon dataset

Why would we want to do all of this?

- Y_1 is known as the **first principal component** in our analysis
- Notice it contains information on all of our original variables X_1 through to X_7
- Y_2 is the **second principal component**, and it too contains information on all of our original variables
- Y_3 is the **third principal component**, and so on and so forth

Each of our principal components is a function of all of our original X variables, and so each one will **capture the overall variability** of our multivariate dataset.

PCA of the heptathlon dataset

The mathematics underpinning the estimation of the ' a 's is such that

$$\text{Var}(Y_1) > \text{Var}(Y_2) > \text{Var}(Y_3) > \dots ,$$

and it's often the case that the first two principal components, considered together, **capture most of the variability** present in our original multivariate dataset.

Thus, our aim is to work with – perhaps – just Y_1 and Y_2 , rather than $X_1, X_2, \dots, X_7$.

Both Y_1 and Y_2 will likely capture most of the variability in our multivariate dataset *and* then we've '**reduced the dimension**' of our analysis from 7-D to just 2-D.

PCA of the heptathlon dataset

- In the jumping events large numbers are indicative of good performances
- In the running events, small numbers are indicative of good performances
- It is good practice, when performing a PCA, to re-scale variables so that the 'direction' is interpreted in the same way for all variables
- Thus, from this point forwards, we now work with a new version of the heptathlon dataset in which the running events have been adjusted so that big numbers indicate good performances
- Specifically, for each athlete j we now have

$$\begin{aligned}X_{1j}^* &= \max(X_1) - X_{1j}, \\X_{4j}^* &= \max(X_4) - X_{4j} \quad \text{and} \\X_{7j}^* &= \max(X_7) - X_{7j}.\end{aligned}$$

PCA of the heptathlon dataset

	Athlete	H100M_scaled	HJ	SP	R200M_scaled	LJ	JT	R800M_scaled
1	Jessica Enni...	2.04	1.86	14.28	3.08	6.48	47.49	31.9
2	Lilli Schwar...	1.32	1.83	14.77	1.14	6.3	51.73	30.05
3	Austra Skuj...	0.58	1.92	17.31	0.48	6.25	51.13	19.96
4	Antoinette ...	1.62	1.8	14.26	1.19	6.13	55.87	26.93
5	Jessica Zeli...	1.93	1.68	14.81	2.59	5.91	45.75	31.4
6	Kristina Sav...	1.21	1.83	14.77	1.45	6.21	43.7	28.28
7	Laura Ikaun...	0.87	1.83	12.64	1.75	6.13	51.27	28.42
8	Hanna Mel...	1.26	1.8	12.96	1.82	6.4	43.86	27.65
9	Brianne The...	1.28	1.83	12.89	1.56	6.01	46.47	31.28
10	Dafne Schip...	1.1	1.8	13.67	3.08	6.28	36.63	25.03
11	Nadine Bro...	0.94	1.86	13.57	0.78	5.94	51.98	23.57
12	Jessica Sam...	1	1.77	14.18	1.66	6.18	42.02	29.24
13	Katarina Jo...	1.1	1.89	11.32	2.18	6.19	38.37	29.79
14	Sharon Day...	1.01	1.77	14.28	1.55	5.85	43.9	29.24
15	Yana Maksi...	0.61	1.89	14.09	0.48	5.99	42.33	27.18
16	Eliska Kluci...	0.57	1.8	12.93	0.91	6.13	45.65	24.47
17	Ellen Sprun...	1.23	1.71	12.62	2.32	5.88	45.63	23.01
18	Olga Kurba...	1.12	1.8	13.71	2.03	5.83	40.36	20.73
19	Marisa De ...	0.84	1.71	13.09	0.65	5.76	51.98	24.35
20	Gyorgyi Far...	0.25	1.8	13.55	0.19	6.07	46.52	22.72
21	Grit Sadeik...	1.08	1.74	12.43	1.66	6.11	44.12	17.54
22	Sofia Ifadid...	0.76	1.68	12.96	0	5.81	56.96	18.52
23	Ivona Dadic...	0	1.8	12.19	1.62	6	41.82	24.65
24	Sarah Cowl...	0.62	1.8	12.37	0.31	6	41.9	21.54

Interpretation of computer output

Principal Components: on Correlations							
Summary Plots							
Loading Matrix							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
H100M_scaled	0.71667	0.22720	0.58599	-0.08939	-0.14086	0.06295	-0.24411
HJ	0.57233	0.20640	-0.70612	0.11652	-0.14599	0.30679	-0.04701
SP	-0.00222	0.76391	0.04788	0.61007	-0.11347	-0.16638	0.03740
R200M_scaled	0.75808	-0.33144	0.42566	0.22369	-0.05684	0.15426	0.23921
LJ	0.78412	0.02781	-0.27212	-0.34724	-0.29418	-0.31262	0.07409
JT	-0.07072	0.83412	0.15170	-0.46408	0.15653	0.13265	0.13699
R800M_scaled	0.75237	0.01219	-0.15191	0.09119	0.62347	-0.11097	-0.03698

Interpretation of computer output

From this output:

$$Y_1 = 0.717X_1^* + 0.572X_2 - 0.002X_3 + 0.758X_4^* + 0.784X_5 - 0.071X_6 + 0.752X_7^*,$$

$$Y_2 = 0.227X_1^* + 0.206X_2 + 0.764X_3 - 0.331X_4^* + 0.028X_5 + 0.834X_6 + 0.012X_7^*,$$

$$Y_3 = 0.586X_1^* - 0.706X_2 + 0.048X_3 + 0.426X_4^* - 0.272X_5 + 0.152X_6 - 0.152X_7^*,$$

But what do these components tell us?

Interpretation of computer output

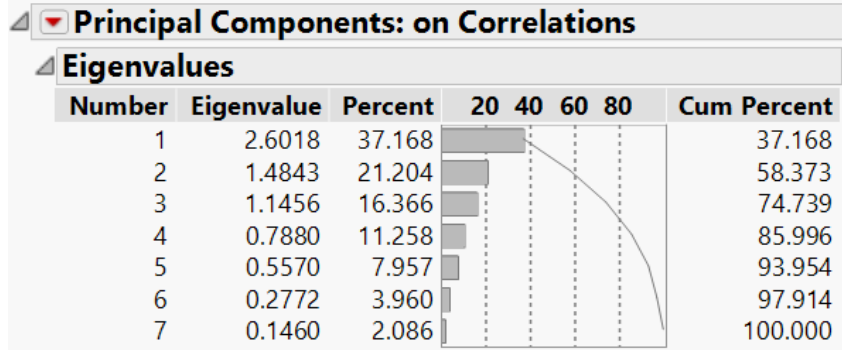
- Each component is a function of all of our original X variables, and in each component the X variables have a 'coefficient'
- Example: In the **first principal component**, the coefficient for the 100 metres hurdles (X_1) is 0.717 and the coefficient for the shot put (X_3) is -0.002
- **First principal component:**
 - The coefficients for the throwing events are both very small relative to the coefficients for the running/jumping events
 - The coefficients of the running and jumping events are all very similar
 - Therefore this principal component might be interpreted as an **average measure of ability in the running and jumping events**

Interpretation of computer output

- **Second principal component:**

- In terms of the coefficients, it is clear that the throwing events dominate this principal component, both having similar coefficients that are larger than the others (relatively speaking)
- Therefore, this principal component might be interpreted as an **average measure of ability in the throwing events**

Interpretation of computer output



Interpretation of computer output

The output on the last slide tells us how much of the variability in our original multivariate dataset is captured by our first three principal components:

- First PC (Y_1): 37.168%
- Second PC (Y_2): 21.204%
- Third PC (Y_3): 16.366%
- So 74.739% of the total variation in our original dataset being captured by the first three principal components. Pretty good!

Calculation of principal component scores

We can now calculate principal component **scores** for each athlete.

Recall that we have:

$$Y_1 = 0.717X_1^* + 0.572X_2 - 0.002X_3 + 0.758X_4^* + 0.784X_5 - 0.071X_6 + 0.752X_7^*$$

$$Y_2 = 0.227X_1^* + 0.206X_2 + 0.764X_3 - 0.331X_4^* + 0.028X_5 + 0.834X_6 + 0.012X_7^*$$

	Athlete	H100M_scaled	HJ	SP	R200M_scaled	LJ	JT	R800M_scaled	Column 9	Column 10
1	Jessica Enni...	2.04	1.86	14.28	3.08	6.48	47.49	31.9	30.53001	50.90758
2	Lilli Schwar...	1.32	1.83	14.77	1.14	6.3	51.73	30.05	26.69175	55.26338
3	Austra Skuj...	0.58	1.92	17.31	0.48	6.25	51.13	19.96	18.12301	56.65008
4	Antoinette ...	1.62	1.8	14.26	1.19	6.13	55.87	26.93	24.15515	58.32967
5	Jessica Zeli...	1.93	1.68	14.81	2.59	5.91	45.75	31.4	29.27636	49.93952
6	Kristina Sav...	1.21	1.83	14.77	1.45	6.21	43.7	28.28	26.01639	48.41502
7	Laura Ikaun...	0.87	1.83	12.64	1.75	6.13	51.27	28.42	25.50936	52.92404
8	Hanna Mel...	1.26	1.8	12.96	1.82	6.4	43.86	27.65	25.983	47.04608
9	Brianne The...	1.28	1.83	12.89	1.56	6.01	46.47	31.28	28.05625	49.29876
10	Dafne Schip...	1.1	1.8	13.67	3.08	6.28	36.63	25.03	25.27095	41.07052
11	Nadine Bro...	0.94	1.86	13.57	0.78	5.94	51.98	23.57	20.99302	54.50632
12	Jessica Sam...	1	1.77	14.18	1.66	6.18	42.02	29.24	26.80954	46.44428
13	Katarina Jo...	1.1	1.89	11.32	2.18	6.19	38.37	29.79	28.03035	41.09732
14	Sharon Day...	1.01	1.77	14.28	1.55	5.85	43.9	29.24	26.34093	48.11804
15	Yana Maksi...	0.61	1.89	14.09	0.48	5.99	42.33	27.18	23.9842	46.93079
16	Eliska Kluci...	0.57	1.8	12.93	0.91	6.13	45.65	24.47	22.06842	48.61488
17	Ellen Sprun...	1.23	1.71	12.62	2.32	5.88	45.63	23.01	22.26706	48.00141
18	Olga Kurba...	1.12	1.8	13.71	2.03	5.83	40.36	20.73	20.63808	44.49979
19	Marisa De ...	0.84	1.71	13.09	0.65	5.76	51.98	24.35	21.18338	54.13335
20	Gyorgyi Far...	0.25	1.8	13.55	0.19	6.07	46.52	22.72	19.86717	49.95714
21	Grit Sadeik...	1.08	1.74	12.43	1.66	6.11	44.12	17.54	17.85086	46.7283
22	Sofia Ifadid...	0.76	1.68	12.96	0	5.81	56.96	18.52	15.91788	58.3096
23	Ivona Dadic...	0	1.8	12.19	1.62	6	41.82	24.65	22.50476	44.48942

Calculation of principal component scores

It can be useful to plot the scores for each athlete on the first principal component against those on the second principal component.

This often reveals **groupings** or **outliers** in our dataset.

Calculation of principal component scores

However, the scores calculated for Jessica Ennis, for example, will be hugely influenced by the size of observations on different variables.

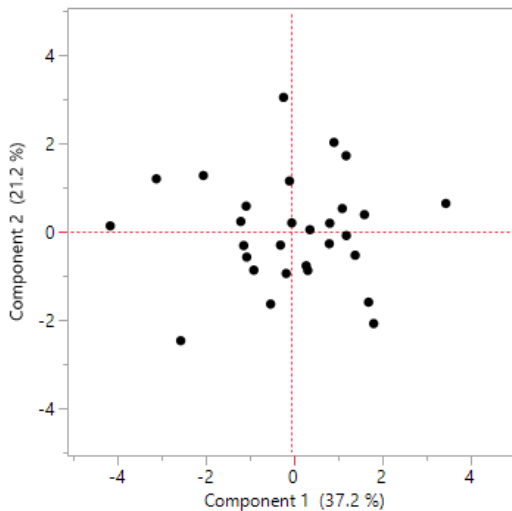
- We could have easily recorded times in the 800 metres running event in milliseconds, giving us even larger values on variable X_7
- If a certain athlete performed particularly well in the 800 metres, but not so well in all the other events, her principal component scores may still be very high

Calculation of principal component scores

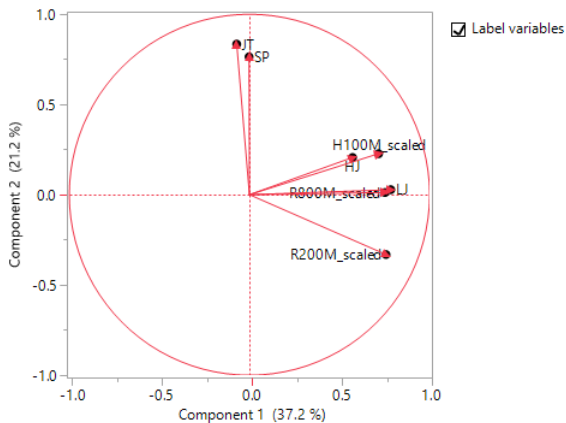
Thus, most standard computer packages calculate principal component scores on the **standardised** outcomes for each variable.

Basically, we subtract the mean and divide by the standard deviation for performances on each variable, and uses these **standardised** data when calculating the principal component scores for each athlete.

Calculation of principal component scores



Calculation of principal component scores



PCA of the microarray dataset

To illustrate the power of PCA in the analysis of 'big data', we now consider a much more detailed dataset!

The **microrarray dataset** consists of measurements on the expression levels of 6,830 genes from 64 samples taken from malignant tumours in different patients.

This really is a **big dataset**!

- Scatterplot matrix: 23 million plots
- No simple way of visualising these data

PCA of the microarray dataset

