Seven into two: Principal components analysis and the Olympic heptathlon

Tom Fanshawe looks at the statistical theory that judges the winner of disciplines that combine several separate elements.

Of all the contests at the Olympic Games, it is surely the multi-event disciplines that provide the most challenging all-round tests of athletic supremacy. But how can we measure who is the most deserving winner when competitors are performing in not just one event, but five, or seven, or ten?

Multi-event competitions at the Olympic Games date to at least the eighth century BC, when the pentathlon was introduced. Mythically the creation of the Greek hero Jason, the original competition consisted of five events: running, standing long jump, discus, javelin and wrestling.

The modern pentathlon, the brainchild of the founder of the Olympic movement, Pierre de Coubertin, stands alongside the men's decathlon and the women's heptathlon as a prominent multi-event discipline in the Games today. Table 1 shows the contributing events for each.

An obvious question arises. In, say, the decathlon, suppose one athlete wins the first five events, and another the second five events; which is the better athlete? Which should be declared the winner? And does the balance of events used favour athletes with particular strengths? One could subdivide the ten events into four that involve jumping in some form – long jump, high jump, pole vault and hurdles – and six that do not. Or one could use the more familiar classification of four track events and six field events. Decisions about the "best" decathlete, heptathlete or pentathlete contain complexities.

Table 1. Composition of multi-event competitions in the Olympic Games

	Decathlon	*Heptathlon*	*Modern pentathlon*
1	100 m	100 m hurdles	Shooting
2	Long jump	High jump	Fencing
3	Shot put	Shot put	Swimming
4	High jump	200 m	Riding
5	400 m	Long jump	Running
6	110 m hurdles	Javelin	
7	Discus	800 m	
8	Pole vault		
9	Javelin		
10	1500 m		

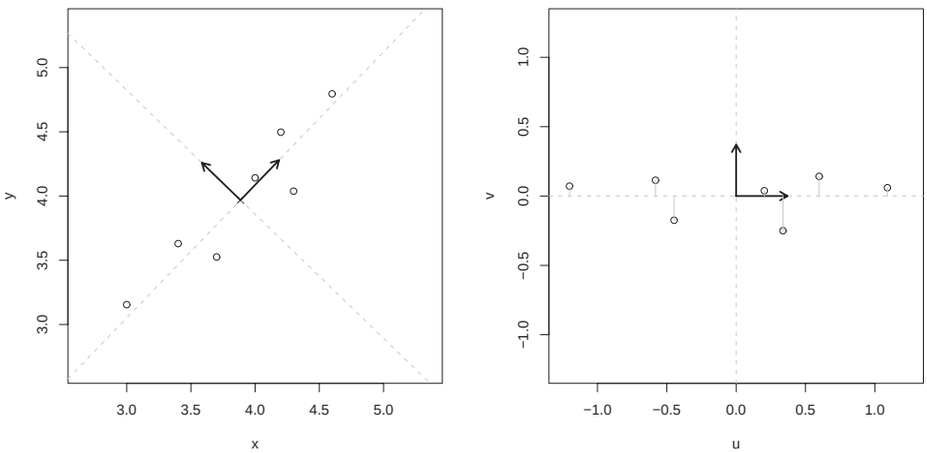

Figure 1. Illustration of dimension reduction by principal components analysis in two dimensions

Principal components analysis is a commonly used statistical technique that enables us to analyse data of just the type that these events produce – called "multivariate data" in statistical jargon. Visualising data in more than two dimensions is difficult, and principal components analysis can help, recognising that variables in multivariate data are often correlated. (The good javelin throwers might very well be the good discus throwers also; the 100m front-runners might also found leading the 400m). Identifying the extent of this correlation enables us to summarise the original data set in a smaller number of dimensions, and allows relationships between the variables to become apparent.

Figure 1 shows a very simple example to illustrate the principal of dimension reduction. The original data, shown in the left-hand panel, are two-dimensional, but clearly highly correlated. Changing the coordinate system from the original (x,y) coordinates to (u,v) coordinates, still orthogonal, and replotting the data produces the rescaled figure in the right-hand panel.

The variance of the u coordinates of the data points is much larger than the variance of the v coordinates. Indeed, much of the variability in the two-dimensional data set can be captured from the u-coordinate value (the "first component") alone, equivalent to projecting each data point onto the u-axis. The essence of principal components analysis is to determine a sequence of orthogonal coordinate axes with respect to which the variance of the data is maximised.

In the heptathlon, data occupy a seven-dimensional space. Performance in each event is summarised by a numerical score, and the athlete with the highest total score across the seven events is the winner. The scoring mechanism thus performs a rather elementary

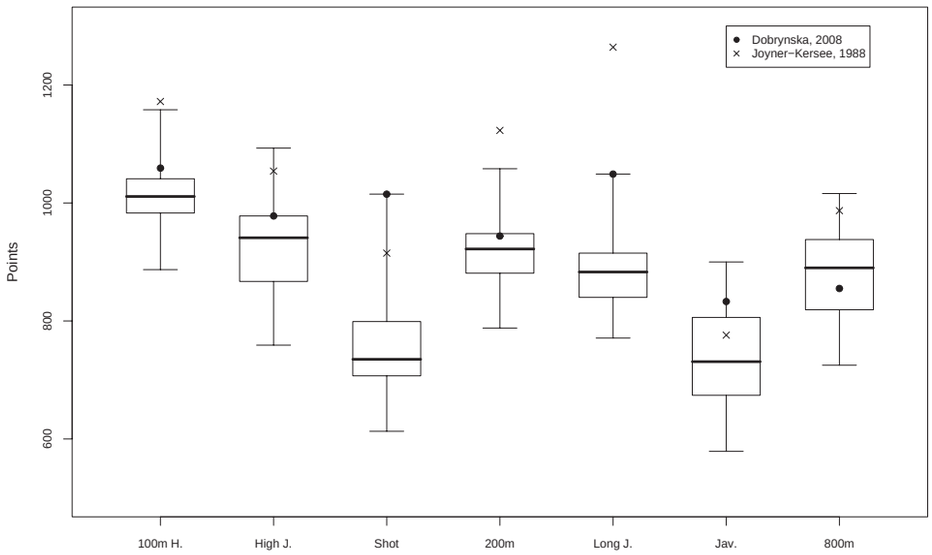

Figure 2. Boxplot of scores for each event in the 2008 heptathlon, with scores from Dobrynska (2008) and Joyner-Kersee (1988) marked. Note that the scoring system for the javelin changed between 1988 and 2008

reduction of the seven-dimensional data into a one-dimensional number, exactly the type of procedure that principal components analysis is able to achieve.

In the women's heptathlon at the 2008 Olympics, 33 athletes completed all seven events (discounting one participant who was disqualified for a doping offence). Ukraine's Nataliya Dobrynska won the gold medal, with 6733 points. The scores of the 33 athletes in each event are summarised in Figure 2, which also shows the scores in each event of both Dobrynska in 2008, and Jackie Joyner-Kersee in the 1988 Olympics, when she set the current world record. Joyner-Kersee's remarkable 1264 points in the long jump is the record for any single event in the women's heptathlon. Athletes tend to achieve lower scores in the two throwing events (shot put and discus) than in the other five.

As might be expected, scores between most events are positively correlated. Figure 3 shows the pairwise sample correlations between scores in each event. The highest correlation, 0.70, is between scores on the high jump and the long jump.

Figure 4, known as a "biplot", summarises the results of a principal components analysis to optimally project the seven-dimensional data set into two dimensions. The resulting two "components" are each made up of linear combinations of the original seven variables, after rescaling, with coefficients given in Table 2. The arrows in Figure 4 represent the seven events, showing the way in which they contribute to each component. Five of the events contribute similarly to the first

Jessica Ennis, heptathlete: if she wins gold in London, will she be the best athlete of the Games? Photo: Craig Brough/Professional Sport. © Professional Sport/TopFoto

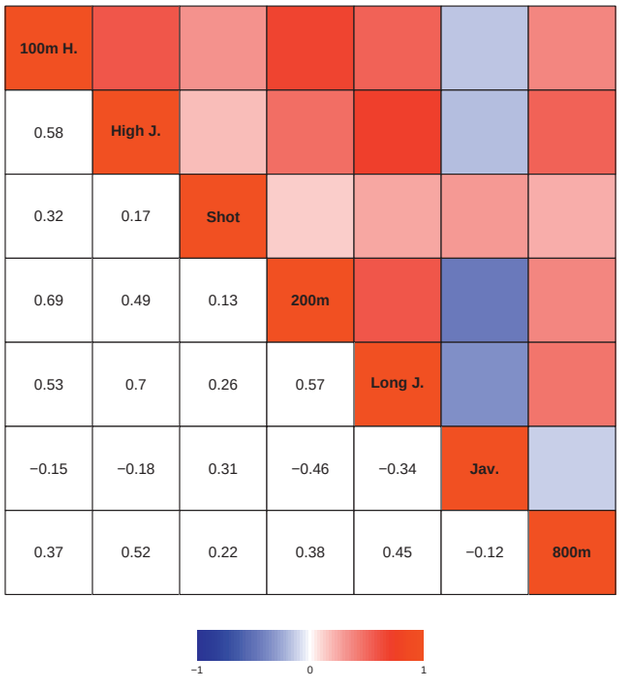

Figure 3. Pairwise correlations between points scored in different events in the 2008 heptathlon. Colours and numbers represent the same quantity: the correlation between scores for events in the corresponding row and column

component, but very little to the second. The other two, the two throwing events, contribute heavily to the second component, but less strongly to the first. The second component is therefore a rough-and-ready measure of proficiency in the throwing events, whereas the first is a broader measure of running and jumping ability.

Notably, scores for the javelin are negatively correlated with scores in all other events except the shot put. Athletes who specialise in throwing events tend to be weaker at running and jumping, and vice versa. This is perhaps not surprising: there are precedents of so-called "runner-jumpers" winning both the 100 metres and the long jump as single disciplines, but few instances of athletes who have successfully added an individual throwing event to their repertoire.

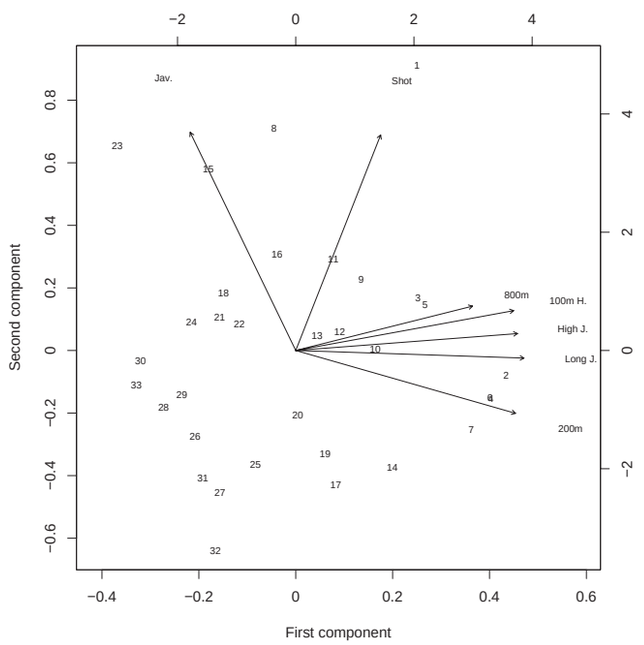

Figure 4. Biplot of principal components analysis of the 2008 heptathlon data. See text for details. Dobrynska, the winner, is represented by number 1

Table 2. First two components from a principal components analysis of the 2008 heptathlon data

	Component 1	*Component 2*
100m hurdles	0.44	0.12
High jump	0.45	–
Shot put	0.17	0.68
200m	0.44	−0.20
Long jump	0.46	–
Javelin	−0.21	0.68
800m	0.36	0.14

The numbers appearing in the background of Figure 4 show the scores from each of the two derived components for the 33 athletes, labelled according to their final position in the competition. The axes on the right and above the plot refer to these scores, each scaled to have zero mean. Of the leading athletes, most score very strongly on the first component, but the winner, Dobrynska, also performs strikingly well on the second. Indeed, it was Dobrynska's performance in the shot, which she launched 17.29 metres, more than 2 metres further than any of her near rivals, that was a key factor in her overall victory.

Currently, the heptathlon contains only two throwing events, although the analysis here confirms the intuition that these events require somewhat different physical strength or skill than those based on running and jumping. The scoring system treats these two events frugally, making high scores in the shot put and javelin difficult to achieve. Yet, as Dobrynska demonstrated, the throwing events are of great importance in deciding the medal positions.

It should be noted that dimension reduction from seven to two dimensions by no means tells the whole story: only two-thirds of the variability in the original scores is explained by the first two components. Interestingly, the first two components from an analysis of modern pentathlon scores from the 2008 Olympics explain only around half of the variability in this five-dimensional data set. In contrast to the heptathlon, the events chosen by Baron de Coubertin appear to test five quite distinct aspects of physical prowess.

This leads us to end on another question: are the heptathlon and decathlon the best measures of all-round prowess in the Olympics? And if so, should we not regard their winners as the best overall athletes of the Games?

Dr Tom Fanshawe is a lecturer in medical statistics at Lancaster University.