Revisiting Olympic Track Records: Some Practical Considerations in the Principal Component Analysis

Dayanand N. NAIK and Ravindra KHATTREE

In some practical problems where a principal component analysis is utilized, the use of the variance covariance matrix of an appropriately defined set of variables, rather than the correlation matrix, may be more meaningful. This is illustrated through the analysis of 1984 Olympic records data on various track events. The analysis results in conclusions that are more appealing to intuition and that are also consistent with a retrospective visual examination of the data on certain leading countries in their athletic excellence.

KEY WORDS: Athletic excellence; Correlation matrix; Principal component analysis; Ranking; Track records; Variance covariance matrix.

1. INTRODUCTION

Principal component analysis is one of the useful techniques of multivariate statistics, being commonly used for the reduction of the dimensionality of datasets. The starting point of a principal component analysis is the spectral decomposition of either a variance covariance matrix, or a correlation matrix, with the objective of identifying only a few but most informative and mutually uncorrelated variables. It is well known that principal component analysis results based on these two matrices can be quite different (Johnson and Wichern 1992). However, in the standard textbooks on multivariate analysis no clear guidelines on which of the matrices should be used are provided. The objective of this short communication is to provide some guidelines for this choice, and also to provide some cautions and considerations to help one decide on the appropriate choice of variables for the analysis. We show that for ranking nations based on their athletic excellence in 1984 Olympic track events, intuitively appealing and more meaningful results can be obtained by using the variance covariance matrix (instead of the correlation matrix) of variables that are defined based on certain physical considerations. For illustration we restrict ourselves to the Olympic track records data given in Dawkins (1989) as the focal point of discussion.

The Olympic track records dataset obtained from Belcham and Hymans (1984) was first analyzed using principal components by Dawkins (1989), with the goal of ranking nations based on their athletic excellence in track events. These data are also provided in Johnson and Wichern (1992), and can be successfully used to illustrate various aspects of principal component analysis in the classroom. The data consist of 1984 Olympic track records of 55 nations for women as well as men. The data matrix for women is a 55×7 matrix with seven events represented, these being the 100 meters, 200 meters, 400 meters, 800 meters, 1,500 meters, 3,000 meters, and marathon (which is 26.2 miles or 42,195 meters long). For the men the corresponding matrix is of order 55×8, differing from the women's events in that the 3,000 meters was excluded, but the 5,000 meters and 10,000 meters were included.

Dayanand N. Naik is Associate Professor, Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077. Ravindra Khattree is Professor, Department of Mathematical Sciences, Oakland University, Rochester, MI 48309-4401. The authors thank Professors J. P. Morgan and G. G. Hegde for their valuable comments and suggestions.

2. THE ANALYSIS

Dawkins (1989) chose to first rescale the variables in each of the two data sets to have mean 0 and standard deviation 1, making them unit free, on the grounds that all the variables are equally important, and hence should somehow be brought to an equal footing. This, in turn, amounts to using the spectral decomposition of the sample correlation matrix instead of the sample variance covariance matrix (of the time taken by the athletes in the events) to obtain the principal components. His objection to the sample variance covariance matrix as a choice for the analysis was also based on the argument that if the raw data were to be analyzed using the same time unit, the variable represented by the time taken in the marathon, due to its larger amount of variability, would be weighted excessively in the analysis.

Dawkins's objections are well taken, and one would readily agree with his concerns. However, his choice of the correlation matrix as the focal point of analysis is suspect in that by forcing all the track records variables to have equal variance by such scaling, the purpose of partitioning the total variability and perhaps the very objective of identifying those variables that contribute more significantly to the total variability have been defeated.

How do we bring all the variables to an "equal footing" while still admitting the possibility that more variability across the nations may be found in certain specific track events? For this, one must obviously use the variance covariance matrix as the basic object for the analysis, but it should correspond to variables that represent characteristics common to all of the events. "Total time taken" is certainly not such a variable. To compare the athletic performances of nations, the appropriate variables should be rate or speed related rather than the total time taken. A variable that may be more relevant in this context is the speed itself, defined as the "distance covered per unit time." This variable succeeds in retaining the possibility of having different degrees of variability in different variables. We will therefore use

Table 1. Speeds for the Track Events (Women)

Country	100 m	200 m	400 m	800 m	1,500 m	3,000 m	Marathon
Argentina	8.61	8.72	7.34	6.20	5.64	5.11	3.94
Australia	8.93	8.95	7.83	6.73	6.05	5.51	4.62
Austria	8.75	8.66	7.90	6.70	5.92	5.35	4.41
Belgium	8.76	8.68	7.69	6.67	6.04	5.63	4.46
Bermuda	8.73	8.68	7.50	6.17	5.46	5.10	4.14
Brazil	8.84	8.63	7.58	6.35	5.57	5.12	4.17
Burma	8.24	8.17	7.27	6.12	5.62	5.26	3.68
Canada	9.09	8.99	7.99	6.67	6.16	5.68	4.71
Chile	8.33	8.16	7.29	6.50	5.91	5.34	4.10
China	8.37	8.19	7.28	6.41	5.77	5.37	4.17
Columbia	8.62	8.33	7.51	6.32	5.75	5.29	4.25
Cook Island	7.75	7.38	6.62	5.80	5.17	4.50	3.02
Costa Rica	8.36	8.13	6.87	6.03	5.34	4.79	4.09
Czechoslovakia	9.02	9.10	8.34	7.05	6.04	5.61	4.43
Denmark	8.76	8.50	7.46	6.57	5.98	5.74	4.63
Dominican Rep.	8.48	8.32	7.14	5.95	5.27	5.06	3.45
Finland	8.98	8.93	7.98	6.57	6.10	5.61	4.56
France	8.97	8.85	7.73	6.67	6.04	5.57	4.53
GDR	9.25	9.21	8.31	6.91	6.31	5.71	4.46
FRG	9.08	8.93	8.04	6.84	6.20	5.82	4.73
GB & NI	9.09	9.04	7.93	6.73	6.20	5.80	4.70
Greece	8.48	8.31	7.28	6.44	5.75	5.07	3.86
Guatemala	8.45	8.15	7.13	5.85	5.14	4.74	3.27
Hungary	8.73	8.67	7.77	6.63	6.04	5.57	4.50
India	8.37	8.24	7.46	6.35	5.79	5.01	3.74
Indonesia	8.44	8.25	7.23	6.01	5.42	4.99	3.49
Ireland	8.75	8.51	7.51	6.50	6.08	5.62	4.71
Israel	8.73	8.49	7.29	6.35	5.88	5.34	4.38
Italy	8.86	8.70	7.69	6.80	6.28	5.79	4.63
Japan	8.53	8.33	7.44	6.38	5.75	5.43	4.67
Kenya	8.53	8.38	7.59	6.67	6.02	5.43	3.88
Korea	8.36	8.17	7.18	6.20	5.66	5.20	4.27
DPRKorea	8.16	7.76	7.81	6.77	5.88	5.35	3.93
Luxembourg	8.31	8.01	7.13	6.44	5.71	5.19	4.03
Malaysia	8.18	8.26	7.26	6.09	5.33	4.78	3.86
Mauritius	8.50	7.97	6.88	5.87	5.22	4.59	2.69
Mexico	8.41	8.47	7.44	6.54	5.88	5.21	4.44
Netherlands	8.89	8.77	7.64	6.70	6.16	5.55	4.61
New Zealand	8.66	8.65	7.75	6.60	5.98	5.71	4.83
Norway	8.64	8.58	7.53	6.57	6.23	5.86	4.83
Guinea	8.16	7.98	7.02	5.95	5.17	4.68	3.02
Philippines	8.50	8.50	7.33	6.09	5.43	4.92	3.51
Poland	8.98	9.00	8.12	6.84	6.27	5.57	4.37
Portugal	8.47	8.26	7.37	6.38	6.01	5.66	4.65
Romania	8.74	8.53	7.81	6.94	6.31	5.86	4.25
Singapore	8.13	8.00	7.26	6.29	5.53	5.03	3.85
Spain	8.47	8.34	7.46	6.50	6.04	5.54	4.33
Sweden	8.96	8.76	7.72	6.60	6.07	5.66	4.55
Switzerland	8.73	8.58	7.53	6.60	6.14	5.70	4.58
Taipei	8.91	8.84	7.62	6.35	5.71	5.19	3.95
Thailand	8.51	8.18	7.17	6.06	5.30	4.86	4.17
Turkey	8.35	8.18	7.09	6.20	5.72	5.33	3.50
USA	9.27	9.16	7.90	6.80	6.33	5.88	4.93
USSR	9.04	9.01	8.13	7.05	6.46	5.92	4.65
Western Samoa	7.85	7.74	6.81	5.72	4.30	3.83	2.30

the speeds in the track events as the variables for the principal component analysis. The separate analyses will be performed for the data sets on women and men, respectively. These two sets (rounded to two digits after the decimal) are presented in Tables 1 and 2, respectively. The corresponding raw data on total time taken are available in Dawkins.

The principal components are calculated using procedure PRINCOMP of the SAS software. As mentioned earlier, the sample variance covariance matrix of the variables, distances (in meters) covered per second for the various track events, and not the correlation matrix, is used for both data sets. Table 3 presents the coefficients of various track events in the first two principal components. Also presented in the same table are the coefficients of the first two principal components extracted from the sample correlation matrix of the original raw data.

Some interesting features of the analysis are worth observing. Justifiably so, the variables depicting the higher levels of variability have been assigned the larger weights in the first principal component for both datasets. As in

Table 2. Speeds for the Track Events (Men)

Country	100 m	200 m	400 m	800 m	1,500 m	5,000 m	10,000 m	Marathon
Argentina	9.62	9.61	8.54	7.37	6.76	5.94	5.68	5.11
Australia	9.70	9.97	8.92	7.66	7.00	6.28	6.03	5.48
Austria	9.58	9.61	8.54	7.45	6.94	6.28	6.01	5.17
Belgium	9.67	9.67	8.88	7.71	6.94	6.30	6.07	5.41
Bermuda	9.73	9.72	8.71	7.41	6.67	5.68	5.46	4.80
Brazil	9.78	9.79	8.85	7.71	6.83	6.12	5.82	5.28
Burma	9.40	9.29	8.28	7.41	6.49	5.77	5.50	5.03
Canada	9.83	9.89	8.76	7.58	6.89	6.15	5.93	5.40
Chile	9.67	9.62	8.66	7.45	6.74	6.12	5.69	5.25
China	9.51	9.51	8.46	7.37	6.70	6.00	5.72	5.27
Columbia	9.59	9.50	8.68	7.33	6.68	6.18	5.98	5.35
Cook Island	8.21	8.62	7.56	6.60	5.90	4.99	4.71	4.27
Costa Rica	9.14	9.13	8.22	7.13	6.51	5.94	5.79	5.15
Czechoslovakia	9.66	9.69	8.76	7.58	6.98	6.21	5.91	5.24
Denmark	9.47	9.75	8.72	7.49	6.93	6.17	5.93	5.38
Dominican Rep.	9.86	9.69	8.55	7.33	6.54	5.59	5.30	4.56
Finland	9.59	9.67	8.79	7.66	6.93	6.28	6.06	5.37
France	9.89	9.81	8.83	7.71	7.00	6.25	5.96	5.32
GDR	9.88	9.84	8.91	7.71	7.02	6.33	6.08	5.41
FRG	9.84	9.82	8.99	7.71	7.08	6.31	6.04	5.32
GB & NI	9.89	9.90	8.90	7.84	7.12	6.41	6.06	5.45
Greece	9.78	9.66	8.59	7.49	6.87	5.71	5.86	5.22
Guatemala	9.11	9.17	8.26	7.05	6.58	5.89	5.54	5.05
Hungary	9.75	9.70	8.69	7.53	6.91	6.18	5.86	5.30
India	9.43	9.34	8.75	7.58	6.70	6.05	5.79	5.33
Indonesia	9.44	9.31	8.37	7.25	6.38	5.66	5.41	4.73
Ireland	9.43	9.54	8.64	7.45	7.02	6.26	5.99	5.31
Israel	9.34	9.52	8.37	7.53	6.72	6.10	5.76	5.11
Italy	9.99	10.14	8.84	7.71	6.94	6.30	6.06	5.37
Japan	9.67	9.61	8.72	7.45	6.87	6.21	6.01	5.47
Kenya	9.56	9.68	8.90	7.71	7.04	6.36	6.09	5.42
Korea	9.67	9.57	8.53	7.45	6.63	5.97	5.70	5.16
DPRKorea	9.17	9.12	8.46	7.21	6.63	5.90	5.62	5.37
Luxembourg	9.66	9.63	8.44	7.33	6.81	6.11	5.73	4.98
Malaysia	9.62	9.56	8.64	7.33	6.58	5.69	5.37	4.56
Mauritius	8.94	8.91	8.39	7.09	6.53	5.53	5.25	4.62
Mexico	9.60	9.39	8.68	7.41	6.85	6.19	5.96	5.44
Netherlands	9.51	9.55	8.87	7.66	6.91	6.24	6.04	5.45
New Zealand	9.51	9.58	8.68	7.66	7.06	6.31	6.02	5.45
Norway	9.48	9.45	8.56	7.58	6.91	6.25	6.02	5.35
Guinea	9.12	9.18	8.35	7.02	6.23	5.66	5.31	4.74
Philippines	9.28	9.24	8.65	7.37	6.53	5.65	5.44	4.84
Poland	9.84	9.88	8.82	7.58	6.94	6.27	5.98	5.34
Portugal	9.50	9.45	8.57	7.45	6.91	6.35	6.09	5.47
Romania	9.61	9.53	8.72	7.58	6.87	6.29	6.02	5.31
Singapore	9.63	9.40	8.44	7.09	6.43	5.52	5.32	4.46
Spain	9.60	9.63	8.70	7.58	7.04	6.26	6.01	5.35
Sweden	9.76	9.70	8.77	7.53	6.93	6.27	5.97	5.38
Switzerland	9.64	9.78	8.74	7.49	7.04	6.30	5.97	5.36
Taipei	9.44	9.39	8.55	7.45	6.63	5.92	5.54	5.05
Thailand	9.62	9.48	8.35	7.29	6.51	5.47	5.12	4.69
Turkey	9.34	9.33	8.40	7.45	6.81	6.15	5.83	5.35
USA	10.07	10.13	9.12	7.71	7.08	6.31	6.08	5.48
USSR	9.93	10.00	8.97	7.62	6.96	6.31	6.05	5.39
Western Samoa	9.24	9.15	8.16	6.60	5.90	5.12	4.80	4.35

Dawkins, here also the first principal component measures the general athletic excellence of a given nation. This is not surprising because here as well as in Dawkins's analysis the corresponding variance covariance and correlation matrices have all nonnegative entries. Hence by the Perron–Frobenius Theorem, all of the coefficients in the first principal component will have the same sign. Assuming that all signs are positive, the first principal component will then represent a weighted average of all the speeds in the various track events. Thus larger scores on the first principal component correspond to higher levels of athletic excellence. The second principal component also has (as in Dawkins) an interpretation as a measure of differential achievement. For both data sets the coefficients in the respective second principal components of the short races are positive and those for long races are negative. The medium distance races have been assigned practically negligible coefficients—a feature not so pronounced in Dawkins's analysis. Nonetheless, the percentages of total variability explained by the first principal component and the first and second components cu-

*Table 3. Coefficients in the First Two Principal Components

Men				Women			
Ours		Dawkins		Ours		Dawkins	
First PC	Second PC	First PC	Second PC	First PC	Second PC	First PC	Second PC
.32	.60	.32	.57	.29	.43	.37	.49
.32	.47	.34	.46	.34	.56	.37	.54
.31	.23	.36	.25	.34	.38	.38	.25
.31	.06	.37	.01	.31	.01	.38	−.16
.34	−.08	.37	−.14	.39	−.20	.39	−.36
.41	−.30	.36	−.31	.40	−.25	.39	−.35
.41	−.30	.37	−.31	.53	−.51	.37	−.37
.38	−.42	.34	−.44				

mulatively remain comparable to Dawkins's for both data sets.

The top 10 nations, based on their athletic excellence as represented by their scores on the first principal component, are listed in Table 4. For comparison we also provide the rankings based on the first principal component extracted from the corresponding sample correlation matrices. These rankings were also reported in Tables 1 and 3 of Dawkins. It is observed that the nations in the top 10 lists, for men as well as women, are the same as those given by Dawkins. For men's track events, the rankings are essentially the same, except that Kenya now outranks France, and their previously assigned ranks as ninth and eighth are now switched. A visual examination of the men's raw data on these two countries (presented in Table 5) puts more intuitive confidence in our ranking in that France was able to outdo Kenya in only two races of relatively short lengths (namely, 100 and 200 meters) out of a total of eight. This observation also seems to question the validity of the promise that the correlation matrix assigns all of the track events equal importance.

The differences in our rankings of nations and that by Dawkins are more in contrast for the women's data. Although the top ten list is still the same, the United States, earlier ranked as third based on the principal component analysis of the correlation matrix, now emerges as the winner! The GDR, on the other hand, is now ranked as third, while the previous analysis declares it as number one. May we add with a bit of humor and sporting spirit, for our German readers, that although unintentionally, we have now set the (track) records straight! Similar shuffling of the countries is observed down the list. For example, countries with previous ranks of 4, 5, and 6 have interchanged their places among themselves; countries with earlier ranks 7 and 8 have interchanged their rankings, and the same is observed for the ninth and tenth ranks.

Table 4. Rankings of Top 10 Nations

Men			Women		
Ours		Dawkins	Ours		Dawkins
1	USA	1	1	USA	3
2	GB & NI	2	2	USSR	2
3	Italy	3	3	GDR	1
4	USSR	4	4	FRG	5
5	GDR	5	5	GB & NI	6
6	FRG	6	6	Czech	4
7	Australia	7	7	Canada	8
8	Kenya	9	8	Poland	7
9	France	8	9	Italy	10
10	Belgium	10	10	Finland	9

Table 5. Men's Track Record Data: Time Taken by Eighth and Ninth Rank Holders*

Track event	France	Kenya
100 m	10.11	10.46
200 m	20.38	20.66
800 m	1.73	1.73
1,500 m	3.57	3.55
5,000 m	13.34	13.10
10,000 m	27.97	27.38
Marathon	132.30	129.75

* The first three times are in seconds, and the remaining in minutes.

A visual examination of data on GDR and the United States is worth pursuing. It is presented in Table 6. The United States is ahead in four out of seven races. Among these four, three are the long distance races, and for the longest two (3,000 meters and marathon), the considerable superiority of the United States over GDR is unquestionable. However, this information is lost as soon as the data are scaled to have unit standard deviations for all variables because the considerable amount of variability across nations in the last two races, especially the marathon, has been wrongly removed, thereby defeating the basic objective of the principal component analysis.

It is worth pointing out that the nations earlier ranked as fourth–sixth based on the first principal component extracted from the correlation matrix, namely Czechoslovakia (Czech), France, and Great Britain and Northern Ireland (GB & NI), seem to follow a more natural ranking when the first principal component is extracted from the sample variance covariance matrix. Heuristically speaking, France outperforms Czech as well as GB & NI by 4:3:0 and 4:2:1, respectively (the third number indicates the number of ties). Among GB & NI and Czech the overall performance results in 4:3:0. This intuitively indicates that among the three, a more appropriate ranking may be France followed by GB & NI followed by Czech. This is precisely the ranking one obtains when the variance covariance matrix is used. How-

Table 6. Women's Track Record Data: Time Taken by First and Third Rank Holders*

Track event	USA	GDR
100 m	10.81	10.79
200 m	21.71	21.83
400 m	48.16	50.62
800 m	1.93	1.96
1,500 m	3.96	3.95
3,000 m	8.75	8.50
Marathon	157.68	142.72

* The first three times are in seconds, and the remaining in minutes.

Table 7. Values of Sample Coefficients of Skewness for x and y

Men			Women		
Event	x	y	Event	x	y
100 m	−1.83	2.32	100 m	−.25	.47
200 m	−.66	.91	200 m	−.25	.53
400 m	−1.30	1.70	400 m	.06	.20
800 m	−1.48	1.77	800 m	−.27	.47
1,500 m	−1.40	1.70	1,500 m	−1.12	1.86
5,000 m	−1.35	1.62	3,000 m	−1.04	1.80
10,000 m	−1.37	1.69	Marathon	−1.22	2.30
Marathon	−1.32	1.51			

ever, the use of the correlation matrix as in Dawkins's results in Czech being declared as the fourth place holder.

As an alternative to speed another variable that may seem relevant is "time taken to cover a unit distance." Between the two variables, time (in seconds) taken to cover 1 meter distance, say y, and the distance (in meters) covered per second, say x, which one is more appropriate? It is well known that the sample variance may not be an adequate representative of the true variance for populations that are skewed. In fact, Rao (1981, p. 124) points out that the "quadratic loss function places undue emphasis on large deviations which may occur with small probability." If the principal component analysis is the potential technique of choice, then the answer would lie in examining the variables y and x for possible symmetry of their distributions. It is advisable to choose that population as the frame of reference that is not overly skewed, for the situation may be especially poor for skewed populations. Because the underlying principle behind a principal component analysis is an appropriate partitioning of the total variance into the variances of several uncorrelated components, we examine the data sets for the variables y as well as x, and choose the one exhibiting less skewness, with the hope that the sample estimates of the variances and covariances from the less skewed population will be better representatives of their population counterparts. An examination of the coefficients of skewness, as well as the box plots and the stem and leaf plots for all track events, suggest that the variable x, the distance (in meters) covered per second, may be the variable of choice for both the data on men and that on women. The sample coefficients of skewness for x as well as y are listed in Table 7 for both datasets and for all track events within the particular dataset. We have suppressed the presentation of all 30 box plots and 30 stem and leaf plots to save space. It may be pointed out that the raw data on the total times taken (and their scaled versions used by Dawkins) would have the same skewness as those for y, and the corresponding plots will also exhibit the same patterns.

One of the referees has pointed out that using the average (or sum) of all seven or eight speed records also results in a ranking very similar to ours. The only differences are the ranks of Kenya and France for men and ranks of GB & NI and Czech for women. Although there possibly is no general theoretical reason for such an agreement, it is not surprising for this data set. The coefficients of all speeds corresponding to this intuitive index (such that the sum of squares of these coefficients is unity) are $7^{-1/2} = .38$ for women's data and $8^{-1/2} = .35$ for men's data. Most of the coefficients obtained by us are close to these values.

3. CONCLUSION

The intent of this article was to discuss certain issues in principal component analysis that are usually not discussed in textbooks. These are illustrated through the 1984 Olympic track events data, for which the consideration of these issues leads to conclusions that are more in line with intuition, especially when the raw data are visually or graphically examined. In essence, we suggest the following. (1) Standardization of the original variables, and hence the use of a correlation matrix, may not always be the correct choice in a principal component analysis. It may destroy the natural and relevant variability present in the data for certain variables. In other words, the fact that the variables are equally important should not be taken to mean that all the variances must be artificially forced to be equal. Further, as we saw in the comparison of the men's data on Kenya and France, use of the correlation matrix does not guarantee that all the variables are indeed being treated equally. (2) Transformations may be used to obtain variables that are more meaningful in measuring the characteristics meant to be measured. For example, judging relative athletic excellence has required variables related to speeds, not the total times taken in the various events (which are quite uneven with respect to distances covered). Hence the extraction of principal components should be based on the variance covariance matrix of these new variables. (3) The variances and total variance are more meaningful indices for measuring variability in data sets that are symmetric. For datasets exhibiting skewness they may not represent the information about variability to its fullest or even adequately. Because a principal component analysis is primarily a partitioning of the total variance, one should prefer variables that exhibit relatively more symmetry over those that are skewed. In the datasets considered here and for all of the track events variables representing distance covered per unit time showed relatively less skewness than those representing the time taken to cover unit distance, and hence the former set of variables is a more appropriate choice.

[Received November 1994. Revised November 1995.]

REFERENCES

Belcham, P., and Hymans, R. (eds.) (1984), *IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympic Games*, London: IAAF.

Dawkins, B. (1989), "Multivariate Analysis of National Track Records," *The American Statistician*, 43, 110–115.

Johnson, R. A., and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, Englewood Cliffs, NJ: Prentice-Hall.

Rao, C. R. (1981), "Some Comments on the Minimum Mean Square Error as a Criterion of Estimation," in *Statistics and Related Topics*, eds. M. Csorgo, D. A. Dawson, J. N. K. Rao, and A. K. Md. E. Saleh, Amsterdam: North-Holland, pp. 123–143.