theoretical point of view, the subject is nicely presented in Muirhead (1982) and Anderson (1984). Principal component analysis was initiated by Pearson (1901) and Hotelling (1933). The asymptotic theory leading to estimates of standard errors was developed by Anderson (1963). Common principal components have been proposed by Flury (1984). A related approach is due to Krzanowski (1979). The 'principal variables' approach to data reduction has been worked out by McCabe (1984).

Software

A good procedure for principal components is contained in SAS, while BMDP and SPSS offer principal component analysis as part of their factor analysis programs. Users who know FORTRAN may prefer a subroutine library like NAG or IMSL for computing eigenvectors or eigenvalues. No software is available, to our knowledge, for computing standard errors. A FORTRAN routine for computing common principal components has been published by Flury and Constantine (1985) and is expected to be released as a subroutine called KPRIN in the IMSL package in 1988.

EXAMPLES

Example 10.1 Decathlon

This example is based on the results of the international decathlon competition at Götzis (Austria), 1982. The data have been collected by Moser (1982), who used face plots to analyse the performance of the best athletes. Parts of Moser's analysis are reported in Flury and Riedwyl (1983).

The raw data, given in Table 10.6, consist of the scores achieved by $n = 20$ athletes in each of the ten categories. The final score assigned to each athlete is the sum of the ten scores, and the rows of Table 10.6 have been arranged in decreasing order of the final scores.

Since the number of observations is small, any multivariate analysis of the data must be exploratory. For the same reason we will analyse only four variables instead of all ten. The four selected variables are the four running disciplines: 100 m dash (DASH), 400 m run (RUN_4), 110 m hurdles (HURDLE), and 1500 m run (RUN_{15}). Table 10.7 displays univariate and bivariate basic statistics for these four variables. The correlations between the first three variables are positive and relatively high, as was to be expected. On the other hand, all correlations with RUN_{15} are small in absolute value, thus confirming the well-known fact that a good short-distance runner need not be a good long-distance runner.

Table 10.8 displays the coefficients of the principal component transformation and the associated eigenvalues. Standard errors of the principal component coefficients computed according to Section 10.7 are given in

Table 10.6 Scores achieved by 20 athletes in the international decathlon competition at Götziz (Austria) 1982. Data courtesy of Urs Moser, University of Berne.

Rank	Name	Nationality	100 m dash	Long jump	Shot put	High jump	400 m run	110 m hurdles	Discus	Pole vault	Javelin	1500 m run
1	THOMPSON	GB	935	1010	807	925	955	926	769	1028	767	585
2	HINGSEN	FRG	817	1004	844	950	905	901	781	957	738	632
3	DEGTJARJOV	USSR	768	893	759	900	825	859	868	981	732	662
4	NIKLAUS	SWITZERLAND	869	867	802	874	915	856	804	884	857	448
5	WENTZ	FRG	787	871	781	874	878	880	782	884	807	592
6	KUELVET	USSR	738	814	700	950	848	850	870	957	764	569
7	STEEN	CANADA	763	887	604	900	862	839	709	1005	753	658
8	BOREHAM	GB	795	853	701	874	890	841	680	859	772	670
9	RUEFENACHT	SWITZERLAND	903	818	700	849	877	919	718	884	716	460
10	KOLOWANOW	USSR	761	846	728	900	765	881	781	981	714	485
11	BAGINSKI	POLAND	747	796	682	849	792	800	746	932	767	564
12	MITRAKIEV	BULGARIA	771	824	668	874	802	840	704	859	710	609
13	HADFIELD	AUSTRALIA	785	911	728	680	878	805	709	884	747	527
14	GUGLER	SWITZERLAND	657	810	698	849	773	820	746	909	771	612
15	ZENIOU	GB	696	774	765	725	785	791	706	932	795	578
16	KUBISZEWSKI	POLAND	724	746	763	849	785	870	724	807	760	509
17	LYTHELL	SWEDEN	712	875	754	725	829	838	762	807	585	516
18	CLAVERIE	FRANCE	756	873	624	725	863	815	655	957	620	474
19	VLASIC	YUGOSLAVIA	622	820	673	769	759	786	698	807	695	619
20	STERRER	AUSTRIA	668	834	601	849	753	751	655	807	642	551

Table 10.7 Basic statistics in the Decathlon example

a) *Means, standard deviations, correlations*

Variable	Mean	Std. deviation	Correlation matrix			
DASH	763.70	77.86	1.00	0.85	0.78	−0.23
RUN_4	836.95	58.54	0.85	1.00	0.61	−0.02
HURDLE	843.40	45.38	0.78	0.61	1.00	−0.10
RUN_{15}	566.00	68.25	−0.23	−0.02	−0.10	1.00

b) *Covariance matrix*

	DASH	RUN_4	HURDLE	RUN_{15}
DASH	6062	3851	2737	−1199
RUN_4	3851	3427	1631	−85
HURDLE	2737	1631	2059	−311
RUN_{15}	−1199	−85	−311	4658

Table 10.8 Principal component transformation in the Decathlon example with four variables. Estimated standard errors of the principal component coefficients are given in brackets

Component	Variable				Eigenvalue
	DASH	RUN_4	HURDLE	RUN_{15}	
U_1	0.756 (0.034)	0.513 (0.082)	0.361 (0.069)	−0.189 (0.266)	10280
U_2	0.052 (0.213)	0.224 (0.168)	0.082 (0.149)	0.970 (0.054)	4548
U_3	0.067 (0.178)	−0.620 (0.151)	0.779 (0.134)	0.073 (0.136)	968
U_4	−0.649 (0.040)	0.551 (0.163)	0.507 (0.202)	−0.135 (0.076)	409

brackets. Of course, we must be very critical of these standard errors – the normality assumptions are highly questionable, and a sample size of 20 is not of a magnitude that guarantees the correctness of asymptotic approximation. Nevertheless, estimated standard errors larger than 0.1 indicate that the corresponding coefficient is very unstable. A quick glance at Table 10.8 suffices to notice that all four components have at least one unstable coefficient. This

Table 10.9 Principal component transformation in the Decathlon example with three variables (estimated standard errors are given in brackets)

Component	Variable			Eigenvalue (variance of U_h)	Std. deviation of U_h
	DASH	RUN$_4$	HURDLE		
U_1	0.762 (0.035)	0.532 (0.060)	0.369 (0.066)	10073	100.4
U_2	0.114 (0.208)	−0.671 (0.167)	0.733 (0.181)	991	31.5
U_3	−0.637 (0.054)	0.517 (0.212)	0.572 (0.230)	483	22.0

is, of course, quite unsatisfactory. We notice, however, that the second component is almost entirely defined by variable RUN$_{15}$, the corresponding coefficient being 0.97. This expresses the fact that RUN$_{15}$ is practically uncorrelated with the other three variables. Hence it may be a good idea to eliminate this variable from the analysis and consider only the remaining three.

Table 10.9 gives the results of a principal component analysis performed on the covariance matrix of DASH, RUN$_4$ and HURDLE. The picture emerging from this analysis is now fairly clear (although the warning regarding the standard errors still applies). The first component seems now sufficiently well defined and can be interpreted as a measure of 'overall performance in short distance running'. However, the three weights (0.762, 0.532 and 0.369) are not equal. The coefficient of DASH is relatively large because this variable has the largest variance among the three variables. The largest eigenvalue (variance of U_1) accounts for about 87% of the total variability, and it is therefore justified to summarize the three-dimensional data by the single principal component U_1.

Let us briefly return to the above interpretation of the first principal component as an overall measure of performance. If we are willing to accept U_1 as a good measure, then we can rank the 20 athletes along this single dimension. The total score (sum of the three or ten disciplines) serves the same purpose. Moreover, as with the first principal component, the total score is a linear combination of the single scores, all coefficients being predetermined and fixed. How do the two methods of ranking compare? In any case, if we are willing to summarize high-dimensional data meaningfully by a final score, then we would expect the final score to be closely related to the first principal component, and the first component should account for a large proportion of the total variability. In our example with three variables, the second condition

Table 10.10 Standard principal components in the Decathlon example with three variables. (*Note: SUM* = sum of scores achieved in three running disciplines. F_1, F_2 and F_3 denote the standardized principal components.)

Athlete no.	SUM	F_1	F_2	F_3	Standard distance
1	2816	2.230	0.027	-0.042	2.231
2	2623	0.977	0.084	1.553	1.837
3	2452	0.027	0.633	0.000	0.634
4	2640	1.260	-0.989	-0.890	1.832
5	2545	0.529	0.062	1.241	1.351
6	2436	-0.112	-0.175	1.176	1.194
7	2464	0.111	-0.639	0.495	0.815
8	2526	0.510	-1.073	0.277	1.220
9	2699	1.548	1.410	-1.129	2.379
10	2407	-0.264	2.399	-0.635	2.495
11	2339	-0.524	-0.113	-1.701	1.784
12	2413	-0.142	0.692	-1.121	1.325
13	2468	0.238	-1.692	-0.651	1.828
14	2250	-1.235	0.432	0.981	1.635
15	2272	-0.982	-0.358	-0.622	1.216
16	2379	-0.479	1.583	0.622	1.766
17	2379	-0.455	-0.143	1.171	1.264
18	2434	-0.025	-1.244	0.097	1.248
19	2167	-1.700	-0.188	0.782	1.881
20	2172	-1.511	-0.708	-1.603	2.314

is fulfilled. To check the first condition, standardized principal component scores were computed according to the formula

$$F_h = U_h/\sqrt{l_h}$$

of Section 10.3. Table 10.10 displays these scores, as well as the sum of the scores achieved in the three disciplines. It is left to the reader to judge how strongly the first principal component and the sum of scores are related, but it is obvious that good runners (i.e. competitors with a high value of F_1) range in general in the upper half of the table, and bad runners in the lower half.

The last column of Table 10.10 gives the standard distance of each athlete from the mean. It is interesting to notice that large standard distances occur at both ends of the list as well as for some 'intermediate' athletes. Decathlonist no. 10 (Kolowanow), for instance, has a standard distance of 2.495. This relatively large value is due to the second principal component, which contrasts the 400 m run with 110 m hurdles. Indeed, Kolowanow's results in these two disciplines do not agree well: below average in the 400 m run, above average in 110 m hurdles. A strong, balanced athlete would be expected to show a high value of the first principal component, and to have scores near 0 in the remaining components. This is indeed the case for the top-ranked athlete no. 1 (Thompson).

Finally, let us add some critical remarks. At the end of Section 10.4 we said that principal component analysis based on the covariance matrix is meaningful only if all variables are measured on the same scale. Do scores achieved in different disciplines have the same unit of measurement? This question is debatable. The same criticism applies to the final score as well – by simply adding up ten scores we implicitly assume that the ten scores are measured on the same scale! Such problems occur quite frequently in the behavioural and social sciences. For instance, measuring the social level or the intelligence of a person depends crucially on the assumption that different scores can simply be added.

Another critical remark refers to the estimation of standard errors. For a sample size of 20 it would be preferable not to use an asymptotic formula. It is indeed possible to estimate standard errors even for small samples by the bootstrap method; see Diaconis and Efron (1983). However, this method is very computer-intensive, and explaining it is beyond the scope of this book.

Example 10.2 Head dimensions of young Swiss men

The data on which this example is based were collected by a group of anthropologists on about 900 Swiss soldiers, most of them recruits. The purpose of the study was to provide sufficient data to construct a new protection mask for the members of the Swiss army. For lack of space we will report here only the data of 200 men, all of them 20 years old at the time of investigation, and only 6 of the 25 head measurements.

In order to explain why principal component analysis was a natural tool in the context of this study, let us give some more details about its purpose. It was clear from earlier investigations that, given the variability in size and shape of human faces, it would be necessary to construct several (two to five) types or sizes of masks to make sure that everybody could be provided with a sufficiently well fitting mask. The anthropologists were therefore asked to identify 'typical' individuals, on whom prototypes of masks would then be modelled.

How can such 'typical individuals' be found? Suppose for the moment that the data were univariate and, say, three types are to be defined. Then one might simply rank all individuals along the single dimension and pick out the ones, for instance, who are at the 15th, 50th and 85th percentiles of the sample distribution. In the multivariate case, an obvious generalization of this idea is to rank all individuals on the first principal component – provided that it is well defined and accounts for a substantial part of the total variance. (Actually the latter proviso is the reason why this simple procedure failed – we will return to this point later.) So the reason for performing principal component analysis on these data was to approximate high-dimensional data by a single variable – the first principal component.

In the actual study, principal component analysis was performed on ten variables that were considered to be important for the fit of protection masks.