Multivariate Analysis: Assignment 2

Michael Gant

03 April 2018

Principle Component Analysis

Underlying Method

Principle Component Analysis (PCA) is a non-parametric method of extracting relevant information from a data set. This is accomplished through decomposing the variance-covariance structure and redundancy reduction.

Decomposing the variance-covariance structure is done via linear combinations of the variables of the observed data $\boldsymbol{X}$. This forms a new, more meaningful basis on which to re-express the data. The new basis can be represented by a simple linear equation:

$$\boldsymbol{AX} = \boldsymbol{Y}$$

where $\boldsymbol{A}$ is the linear transformation matrix and $\boldsymbol{Y}$ is the new representation of the data. This equation gives a basic overview of what PCA does. The question now is what linear combination to choose.

The derivation of PCA is based on finding coefficients that maximise $\boldsymbol{a'\Sigma a}$ subject to $\boldsymbol{a'a} = 1$ where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$. The coefficient vectors $\boldsymbol{a}_i$ define the linear combinations that result in $\boldsymbol{Y}_i$ as follows:

$$\begin{aligned}\boldsymbol{Y}_1 &= \boldsymbol{a}_1'\boldsymbol{X}\\ \boldsymbol{Y}_2 &= \boldsymbol{a}_2'\boldsymbol{X}\\ &\vdots\\ \boldsymbol{Y}_p &= \boldsymbol{a}_p'\boldsymbol{X}\end{aligned}$$

where $\text{Var}(\boldsymbol{Y}_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma a}_i$ and $\text{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma a}_k$ for $i, k = 1, .., p$

It can be shown that the coefficient vectors, $\boldsymbol{a}_i$, are the eigenvectors of the covariance matrix of $\boldsymbol{X}$. The associated eigenvalues, λ_i, show how much of the total variance is decomposed by the corresponding eigenvector with large eigenvalues explaining more variance than smaller ones.

Thus we can find $\boldsymbol{Y}$ as the matrix of row vector principle components using the eigenvectors of $\boldsymbol{\Sigma}$. As eigenvectors are not correlated, the covariance matrix of $\boldsymbol{Y}$ is a diagonal matrix. Thus it can be shown that $\sum_{i=1}^{p} \text{Var}(\boldsymbol{X}_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} \text{Var}(\boldsymbol{Y}_i)$ and we can calculate the total proportion of variance due to the j'th principle component as:

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

##PCA Application to National Track Records

The objective of applying PCA to the given National Track Record data is to determine the linear combinations of variables that explain the majority of the variance. The female record times for 100m, 200m, 400m, 800m, 1500m, 3000m from 55 countries were analysed first. The first 2 principle components are tabulated below with their respective variance proportion:

Table 1: Summary of First 2 Principle Components for Female Event Record Times

	m100	m200	m400	m800	m1500	m3000	Marathon	Eigenvalue	Proportion of Variance
PC1	-0.016	-0.039	-0.108	-0.005	-0.013	-0.039	-0.992	274.366096	0.984
PC2	0.115	0.290	0.938	0.013	0.036	0.079	-0.119	4.016016	0.014

The first principle component combines the variables with most of the weight on the marathon record time variable. This is difficult to interpret as it seems to be an overall athletic record score but with the majority of the emphasis on the marathon record time. The negative weights are present as a lower record time is better and so the centered event time for an above average performance will be negative.

The second principle component has positive coefficients for all different records except the marathon record. Again, this is difficult to interpret as the 400m record coefficient is significantly larger than the rest. The proportion of variance explained by the first principle component is 98.4% and the second is 1.4%. The first principle component explains more than enough of the variance and the second does not contribute much.

Multiplying the data matrix $\boldsymbol{X}$ by the eigenvector related to the first principle component gives a marathon-dominated athletic record score for each country. The rankings of the countries and their respective scores are summarised below:

Table 2: Rankings of Countries by first Principle Component score (Female)

Rank	PC1	Country
1	18.581	GreatBritain
2	15.097	Kenya
3	14.452	China
4	14.113	Japan
5	12.817	U.S.A.
6	12.639	Germany
7	12.616	Russia
8	12.43	Norway
9	11.436	Ireland
10	11.326	Romania

From the rankings it can be seen that Great Britain has the highest score which is due to the country having the fastest female marathon record time.

The data is now transformed from the record time of the events to the record speed in the event measured in meters per second. The prior PCA methods are applied on the converted and scaled data. The data is scaled in order to equalise the deviations in speeds in different events. As the speed differences in shorter races are significantly smaller than those of the long-distance events, not scaling would overlook the short-distance event deviations and focus on the long-distance event speed. Thus, scaling allows more insight of a countries record speed over all the events.

Table 3: Summary of First 2 Prinicple Components for Female Event Record Speeds (M/s)

	m100	m200	m400	m800	m1500	m3000	Marathon	Eigenvalue	Proportion of Variance
PC1	0.374	0.381	0.367	0.393	0.390	0.381	0.359	5.832225	0.833
PC2	-0.425	-0.417	-0.438	0.152	0.282	0.397	0.440	0.648025	0.093

The first 2 principle components can be interpreted much more easily than before. The first principle component represents an overall event record speed score, with all the events contributing similarly to the score. The positive coefficients mean that higher record speeds are better than lower speeds.

The second principle component can be interpreted as a contrast of the short distance events to the long-distance events. PC2 essentially measures what type of events a country performs relatively well in. A positive value would indicate a country who performs better in long-distance events and conversely a negative value would indicate a country who performs better at short distance events. Values close to zero would indicate a country who performs equally in both distance events.

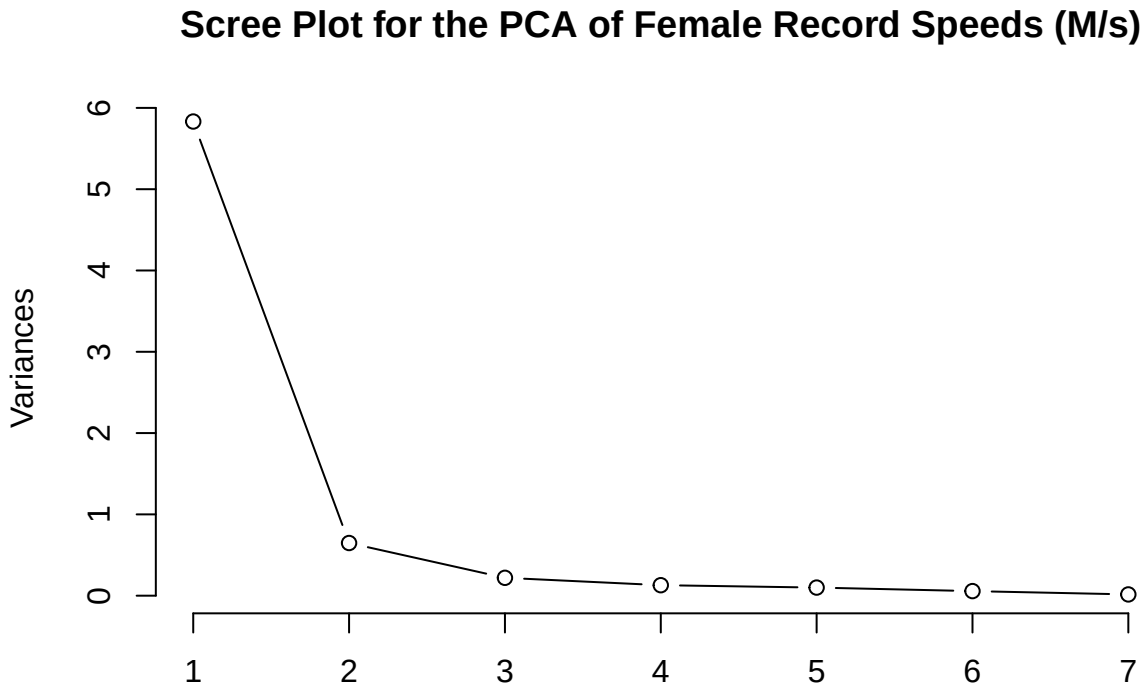
The proportion of variance explained by PC1 and PC2 is 83.3% and 9.3% respectively. The cost of obtaining a principle component that is clearer to interpret is a lower proportion of variance explained. However, the first 2 principle components combined explain over 92% of the variance, more than enough to justify transforming the data to meters per second. The meters per second transformation is also beneficial as speed combines the information from the time taken for the event and the distance of the event.

The rankings of the countries on PC1 have changed and the first 10 are shown below:

Table 4: Rankings of Countries by first Principle Component score (M/s)

Rank	PC1	Country
1	3.598	U.S.A.
2	3.318	Russia
3	3.316	China
4	3.304	Germany
5	2.681	GreatBritain
6	2.639	France
7	2.533	CzechRepublic
8	2.395	Poland
9	2.3	Romania
10	2.019	Australia

Plotting a scree plot below shows that the first 2 principle components are the most important when using the visual elbow rule.



Moving on to the PCA of the male records, the data is slightly different as there are 2 events which males compete in that women don't and 1 event that females compete in that males don't. The following 2 tables show the first 2 principle components for the male events with record times and record speeds respectively:

Table 5: Summary of First 2 Prinicple Components for Male Record Times

	m100	m200	m400	m800	m1500	Marathon	m5000	m10000	Eigenvalue	Proportion of Variance
PC1	-0.017	-0.044	-0.114	-0.005	-0.015	-0.974	-0.079	-0.175	84.511249	0.983
PC2	0.100	0.253	0.916	0.014	0.031	-0.167	0.117	0.209	1.140624	0.013

Table 6: Summary of First 2 Prinicple Components for Male Records Speeds M/s

	m100	m200	m400	m800	m1500	Marathon	m5000	m10000	Eigenvalue	Proportion of Variance
PC1	0.332	0.345	0.337	0.354	0.368	0.354	0.371	0.366	6.625476	0.828
PC2	0.522	0.471	0.361	-0.093	-0.161	-0.380	-0.294	-0.333	0.677329	0.085

The interpretation of the first 2 principle components for the male record time data is exactly the same as for the female record time data. Therefore, it makes sense to do the meters per second transformation. Subsequently, it can be seen that PC1 measures overall event record speed score for each country, with all events being considered. For PC2, the negative and positive coefficients indicate that the principle component contrasts short and long-distance event record speeds, with a slight emphasis on the ultra-long-distance events.

As with the female record speed scores, the rankings of the male record time and speed scores by the first principle

components are summarised in the table below:

Table 7: Rankings of Countries by first Principle Component score (Male)

Rank	PC1	Country
1	9.326	Kenya
2	8.528	U.S.A.
3	7.522	Brazil
4	7.469	Japan
5	7.34	France
6	7.202	Portugal
7	6.646	GreatBritain
8	6.584	Mexico
9	6.576	Belgium
10	6.456	Spain

Table 8: Rankings of Countries by first Principle Component score (M/s)

Rank	PC1	Country
1	4.125	U.S.A.
2	3.085	GreatBritain
3	2.878	Kenya
4	2.536	France
5	2.464	Australia
6	2.345	Italy
7	2.281	Brazil
8	2.239	Germany
9	2.185	Portugal
10	2.098	Belgium

There is a significant change in the rankings after transforming and scaling the data.

Comparisons between the principle component analysis results from the female data to the male data should be approached with caution due to the set of different events that each gender competed in. The values of the coefficients and eigenvalues cannot be compared due to the scaling and centering of the data. However, the proportion of variances can be compared and it can be seen that the proportion of variance explained is very similar across genders.

Factor Analysis

Underlying Method

The primary motivation for Factor Analysis is to describe the covariance relationships among the latent (unobservable) variables. These latent variables are measurements that cannot be directly observed due to their nature, such as happiness or standard of living. Proxies can be used to estimate their values, but it is not possible to get a true measure. These latent variables are called factors. The assumption is made that if there are variables that are highly correlated, they act as proxies and represent the underlying factor. The factor analysis model is mathematically represented as:

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{X}$ is the observed data, $\boldsymbol{\mu}$ is the mean vector of the data, $\boldsymbol{L}$ is the matrix of factor loadings, $\boldsymbol{F}$ is the vector of unobserved common factors and $\boldsymbol{e}$ is the vector of unobserved specific factors. The elements of the factor loading matrix, l_{ij}, represent the estimate of the association between the i'th observed variable and the j'th unobserved factor.

The factor analysis model is similar to a multiple regression model, but the "explanatory variables" are the unobserved factors, $f_1, f_2, ..., f_m$ where m is less than the number of variables p. Thus, the factor analysis model represents the data in a lower dimension.

Under the factor analysis model, the covariance matrix of the data, $\boldsymbol{\Sigma}$, can be expressed as:

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}$$

where $\boldsymbol{\Psi}$ is the diagonal covariance matrix of $\boldsymbol{e}$. The right-hand side is a simplified representation of $\boldsymbol{\Sigma}$ depending on the chosen number of factors, m. Ideally the number of estimated values $mp + p$ is substantially smaller than the number of unique $\boldsymbol{\Sigma}$ estimates, $p(p+1)/2$, without being too small so as not to be able to adequately represent $\boldsymbol{\Sigma}$.

Before going into estimation of the parameters, the rotational ambiguity of factor analysis needs to be addressed. When the number of factors is greater than 1, it can be shown that there are an infinite number of orthogonal matrices that satisfy the required assumptions and equivalently fit the data. This ambiguity is exploited to justify the factor rotation in order to obtain a more parsimonious and interpretable model.

There are 2 different ways to estimate the parameters of a factor model, namely:

- Principle Component method
- Maximum Likelihood Estimation method

The Principle Component method calculates the sample covariance matrix, $\boldsymbol{S}$, and then, using spectral decomposition, the factor loading matrix is estimated by the eigenvalue-eigenvector pairs of $\boldsymbol{S}$. The estimate of the specific variance matrix, $\boldsymbol{\Psi}$, is given by the diagonal elements of $\boldsymbol{S} - \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}'$.

If the variables are standardized, the sample correlation matrix used is $\boldsymbol{R}$. The choice of m is the number of positive eigenvalues of $\boldsymbol{S}$ or the number of eigenvalues of $\boldsymbol{R}$ greater than 1.

The Maximum Likelihood Estimation method makes a normal distribution assumption about the common factors, $\boldsymbol{F}$, and the specific factors, $\boldsymbol{e}$. Thus, maximum likelihood estimates for the factor loadings and the specific variances can be obtained through numerical maximization.

The ambiguity of rotation was discussed earlier and is brought up again. As factor models are equivalent to a rotated factor model, an appropriate rotation may be found that results in more interpretable factors. The Varimax rotation is one such appropriate rotation that maximises the variances of the squares of the factor loadings. Another rotation is the promax method of oblique rotations, which is used when factors are not independent from one another.

Once a satisfactory factor model has been created and rotated, the factor scores can be estimated. Factor scores are the estimated values of the common factors and are similar to principal components. There are 2 methods of estimation that are considered, namely, weighted least squares and regression.

Application

As with PCA in the previous section, the National Track Record data is analysed using factor models. Firstly, the female record speeds are considered. Due to the consideration of only the first 2 principle components from the PCA section, the factor analysis will only assume 2 underlying factors. The maximum likelihood method is used to estimate the loading matrix and is tabulated below:

Table 9: Factor Loadings estimated using Maximum Likelihood Estimation from Female Record Speeds (M/s)

	m100	m200	m400	m800	m1500	m3000	Marathon	Proportion of Variance
Factor 1	0.8804116	0.9109752	0.844602	0.9207781	0.9657499	0.9443679	0.8336690	0.812
Factor 2	0.3471652	0.3911207	0.349982	-0.0440971	-0.1951144	-0.3038984	-0.2048342	0.081

The loadings of factor 1 are all very close to 1, indicating a strong influence on the variables. For the second factor, the majority of the loadings are significant except the loadings for the 800m event.

The loadings are now rotated to obtain better interpretations. As the variables are assumed to not be independent, the oblique rotation promax method is used and the following factor loadings are obtained:

Table 10: Rotated Factor Loadings estimated using Maximum Likelihood Estimation from Female Record Speeds (M/s)

	m100	m200	m400	m800	m1500	m3000	Marathon	Proportion of Variance
Factor 1	0.066	0.022	0.039	**0.653**	**0.899**	**1.041**	**0.828**	0.43
Factor 2	**0.895**	**0.974**	**0.884**	0.319	0.109	-0.065	0.039	0.379

The bold values indicate the factor loadings that have a significant effect on the variable. The rotation separates the significant factor loadings into 2 groups, 4 loadings related to long-distance events in factor 1 and 3 loadings related to short-distance events in factor 2. Thus we can interpret factor 1 as the underlying long-distance performance factor and factor 2 as the short-distance performance factor.

The above analysis is repeated for the male record speeds. The following results are obtained:

Table 11: Factor Loadings estimated using Maximum Likelihood Estimation from Male Record Speeds (M/s)

	m100	m200	m400	m800	m1500	Marathon	m5000	m10000	Proportion of Variance
Factor 1	0.880	0.911	0.845	0.921	0.966	0.944	0.834	0.812	0.792
Factor 2	0.347	0.391	0.350	-0.044	-0.195	-0.304	-0.205	0.081	0.093

Table 12: Rotated Factor Loadings estimated using Maximum Likelihood Estimation from Male Record Speeds (M/s)

	m100	m200	m400	m800	m1500	Marathon	m5000	m10000	Proportion of Variance
Factor 1	-0.035	-0.068	0.205	**0.559**	**0.71**	**0.963**	**0.976**	**1.014**	0.472
Factor 2	**0.954**	**1.031**	**0.7**	0.371	0.261	-0.019	0.02	-0.027	0.334

Canonical Correlation Analysis

Underlying Method

Canonical Correlation Analysis is used to describe the relationships between 2 different sets of variables, $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. This is achieved by analysing the correlation between linear combinations of the variables from the one set to linear combinations of variables from another set. Pairs of linear combinations that have the largest correlation are determined and shown below:

$$U_1 = \boldsymbol{A}_1' \boldsymbol{X}^{(1)} = \sum_i a_{1i} x_i^{(1)}$$
$$V_1 = \boldsymbol{B}_1' \boldsymbol{X}^{(2)} = \sum_i b_{1i} x_i^{(2)}$$

Then the pair of linear combinations with the largest correlation among all pairs uncorrelated with the initial pair is determined. This set can be repeated to get p pairs. The derivation of the canonical covariates shows the following properties:

- $\text{Var}(U_k) = \text{Var}(V_k) = 1$
- $\text{Cov}(U_k, U_l) = \text{Corr}(U_k, U_l) = 0$
- $\text{Cov}(V_k, V_l) = \text{Corr}(V_k, V_l) = 0$
- $\text{Cov}(U_k, V_l) = \text{Corr}(U_k, V_l) = 0$

Application of CCA to Nation Track Record Data