SAGE researchmethods

Principal Components Analysis

Foundation Entries

SAGE Research Methods Foundations

By: Graeme D. Hutcheson
Published:2020
Length: 10,000 Words
DOI: http://dx.doi.org/9781529749335
Methods: Principal Components Analysis
Online ISBN: 9781529749335
Disciplines: Anthropology, Business and Management, Criminology and Criminal Justice, Communication and Media Studies, Counseling and Psychotherapy, Economics, Education, Geography, Health, History, Marketing, Nursing, Political Science and International Relations, Psychology, Social Policy and Public Policy, Social Work, Sociology, Science, Technology, Computer Science, Engineering, Mathematics, Medicine
Access Date: September 4, 2021
Publishing Company: SAGE Publications Ltd
City: London

Abstract

Principal component analysis (PCA) is a technique that essentially converts observed correlated variables into unobserved uncorrelated components. This enables a data set containing many individual variables to be described using a small number of components that capture much of the variation in the data set. PCA has a long history in statistics and has been applied in many disciplines including biology, astronomy, geography, social sciences, meteorology and management. In addition to reducing the number of variables required to describe a data set, PCA can also identify underlying mechanisms that may have played a role in determining the structure in the data (i.e., the underlying “causes”). The reduction of a large number of variables to a relatively small number of components also enables a data set to be more easily analysed and described using other techniques. In particular, as the components identified by PCA are uncorrelated, many of the problems associated with multicollinearity are alleviated, enabling regression models to be more easily interpreted. This entry provides a relatively nontechnical and practical introduction to the application of PCA using a readily available data set and open-source software.

Introduction

The principles behind principal components analysis (PCA) can be illustrated using a simple data set, which contains a number of inter-related variables. Consider the hypothetical correlation matrix in Table 1 which shows the relationships between athletes’ performance in a number of different sporting events.

Table 1. Correlation matrix for associated variables.

	100 m swimming	**400 m hurdles**	**Long jump**	**Shot put**	**Weight lifting**	**Javelin**
100 m swimming	-					
400 m hurdles	**0.80**	-				
Long jump	**0.78**	**0.73**	-			
Shot put	0.22	0.14	0.19	-		
Weight lifting	0.04	0.11	0.13	**0.89**	-	
Javelin	0.24	0.25	0.29	**0.69**	**0.71**	**-**

A visual inspection of Table 1 suggests that performance in 100 m swimming, 400 m hurdles, and the long jump appears to be closely related (as indicated by correlations between 0.73 and 0.8). Similarly,

performances in the shot put, weight lifting, and javelin also appear to be closely related. The six sporting events appear to form two distinct clusters of events. These clusters are not highly related to each other, as indicated by the low correlation coefficients between the variables in each cluster. The sporting events may cluster due to a common "cause" (i.e., there might be an underlying reason why those athletes who are good at shot put are also good at weight lifting). With this example, it is not difficult to hypothesise what this cause may be: performance in swimming, track and long jump require athletes who are fast and lean, whereas shot put, weight lifting, and javelin predominantly require strength. It may be possible to describe and predict athletes' performance using information about their cluster membership (i.e., speed vs. strength) rather than their performance in the six individual events. This potentially reduces the number of pieces of information needed to describe the data (the aim of data reduction) and also gives some insight into the attributes that may be driving sporting performance in individual events. A description of the athletes' performance given in terms of two underlying components as opposed to their performance in six individual and disparate sporting disciplines results in a simpler interpretation of the data, and one that is, perhaps, theoretically more satisfying.

A common example of the use of PCA is in the identification of attitudes through questionnaire research. In a large questionnaire, a number of questions will often address similar issues (e.g., respondents' behaviour, feelings, and beliefs) which will lead to answers that are correlated. For example, answers of "agree" or "strongly agree" to questions such as "It is important to preserve one's culture," "I am prepared to die for my country," and "It is good to take part in our traditional festivals" may lead one to conclude that the respondent is "patriotic." Here, patriotism is not a single measurable entity but a construct which is derived from the measurement of other, directly observable variables (the individual questions). Patriotism explains some of the relationships between the variables and its identification can simplify the description of the data and help in the understanding of complex relationships. Postulating the existence of something called "patriotism" explains to some extent the observed correlations between responses to numerous and varied situations. It is important to note that patriotism here is not a directly observed variable (i.e., there is not a question that asks directly "How patriotic are you"), but is a construct inferred from the relationships in the data.

Each variable in a data set may be expressed as a combination of underlying components (or causes) that are not actually observed. For example, a student's result in an examination might be influenced by a number of factors, such as their aptitude in that particular subject, their experience with taking examinations, their IQ, their writing ability, and even their physical and mental health on the day of the test. A student's score in an examination will not just be a reflection of their ability in that particular subject but will also indicate a number of other abilities or conditions. A student's score, in say a mathematics or English test, may be described in terms of these underlying components using a model which is analogous to a multiple regression. Here, the observed variable (the marks in a particular test) is modelled using a linear combination of components which are not observed but are derived from the data. These equations simply state that a student's score in each observed variable may be predicted from a number of underlying abilities and conditions that are derived from the data. In general terms, a variable may be represented by a number of underlying components:

$$\text{Variable} = a\ \text{Component1} + b\ \text{Component2} + c\ \text{Component3} + \ldots + k\ \text{ComponentN}$$

When this is applied to the simple example of examination scores, individual variables may be represented in terms of a number of underlying abilities or conditions:

$$\text{Mathematics Score} = a\ \text{IQ} + b\ \text{Experience} + c\ \text{Writing ability} + \ldots + k\ \text{Aptitude}$$

$$\text{English Score} = a\ \text{IQ} + b\ \text{Experience} + c\ \text{Writing ability} + \ldots + k\ \text{Aptitude}$$

Mathematics Score and English Score are observed variables and IQ, Writing ability, and Aptitude are unobserved components. The subscripts *a* to *k* indicate the strength of the relationship between the individual components and the variable.

In much the same way as the observed variables can be represented by the underlying unobserved abilities or conditions, these abilities and conditions may also be represented by the observed variables. A linear combination of variables indicates a “pool of correlation” between the observed variables that may indicate a common cause. In general terms, a component may be represented by a linear combination of the observed variables:

$$\text{Component} = a\ \text{Variable1} + b\ \text{Variable2} + c\ \text{Variable3} + \ldots + k\ \text{VariableN}$$

When this is applied to the simple example of examination scores, an individual component may be represented in terms of the observed variables:

$$\text{Component} = a\ \text{Physics} + b\ \text{English} + c\ \text{Geography} + \ldots + k\ \text{Computing}$$

This individual component will account for one underlying source of variation in the data. Additional sources of variation may also exist, which can be identified by analysing the remaining variation in the data (the residuals) after this component has been accounted for. This process continues until all of the variation in the data set is accounted for when the number of components equals the number of variables.

Each component is the linear combination of variables that accounts for the largest amount of the remaining variance in the data set, after accounting for previous components:

$$\text{Component 1} = a\ \text{Physics} + b\ \text{English} + c\ \text{Geography} + \ldots + k\ \text{Computing}$$

$$\text{Component 2} = a\ \text{Physics} + b\ \text{English} + c\ \text{Geography} + \ldots + k\ \text{Computing}$$

$$\text{Component k} = a\ \text{Physics} + b\ \text{English} + c\ \text{Geography} + \ldots + k\ \text{Computing}$$

Component 1 is the linear combination of variables that accounts for the largest amount of variance in the data set. If this component accounts for 40% of the variability in the data, there remains 60% of the variance to still be accounted for. Component 2 is the linear combination of variables that accounts for the largest

amount of variance that is left in the data set (i.e., the 60% of variability that is uncorrelated with Component 1). Successive components account for smaller and smaller proportions of the total variance in the data until all the variance is accounted for (this happens when the number of components is the same as the number of variables). This procedure has converted *k* observed variables into k unobserved components. This is essentially what PCA does: It converts observed correlated variables into unobserved uncorrelated components.

Although a PCA produces components, it does not assign any description (or meaning) to these. The components merely identify the linear combination of variables that explain the variance in the data. Any descriptions of the components and any underlying structure that they reveal is determined by careful interpretation of the results, usually in graphical format. Interpretation of a PCA is not a mechanical process as it involves a number of decisions that cannot be encapsulated in simple rules.

A successful analysis requires the careful selection and screening of variables to ensure that there is an underlying structure to the data and that the individual variables contribute to this. The analyst also needs to consider how many components are needed to adequately represent the data. The number of components selected depends not only on the amount of variance accounted for but also on the aims of the analysis and the likely strength of the underlying structure that is being identified. Although a PCA is, conceptually, a relatively simple analytic technique, it is a creative one and one that requires careful consideration. Perhaps the best way to illustrate PCA is through a worked example.

A Worked Example of PCA

The technique of PCA is demonstrated in this entry using data which shows children's performance on a number of tests (these data are explained in detail in Hutcheson & Sofroniou, 1999). These tests were designed to assess a range of different abilities and skills in order to identify general underlying skills that may underpin children's general development. For example, it is suspected that some children are particularly good at language and communication, whilst others are better at spatial tasks or with social interactions. The PCA will analyse the structure in these data and hopefully provide evidence for the existence of these basic underlying skills.

PCA is a common technique and can be applied using a variety of statistical packages and spreadsheets including R, SAS, Stata, SPSS, S-plus, and Excel. The package used for the worked example here is R (Crawley, 2013; R Core Team, 2018), which was chosen because it is freely available for all computing platforms and includes a number of libraries that provide a comprehensive range of techniques and graphical user interfaces that greatly simplify the data analysis process. The present analysis has made use of the Rcmdr interface (Fox & Bouchet-Valat, 2018) and the FactoMineR plugin (Husson, Josse, & Le, 2016; Le, Josse, & Husson, 2008). The data are available for download in csv and RData formats (PCAexample.csv,PCAexample.RData) as well as the associated R-commands (PCAexample.Rmd) from www.research-training.net.

Table 2 shows the 19 variables which were considered for inclusion in this analysis. The abilities of the children were assessed using 5-point scales and a scoring method that assigned high scores to positive relationships; for example, a higher score for the variable ***Active*** indicates a child who was more active. Similarly, a higher score in ***Newsit*** indicates a child who is more competent in new situations. Data were collected from 293 children and are available for download. This is a relatively small data set which does not contain any missing data or unusual relationships and is designed primarily for demonstration and ease of analysis.

Table 2. Variables in data file.

Variable Label	Variable Description
Active	How active
Artic	Articulation
Atten	Attention
Comp	Comprehension
Coord	Coordination
Day	Day of week born on
Draw	Drawing
LangExp	Expressive language
Math	Mathematical ability
Month	Month born on
Motsk	Motor skills
Newsit	Capability in new situations
SocInt01	Social interaction measure 1
SocInt02	Social interaction measure 2
SentCom	Sentence completion

Variable Label	Variable Description
Temp	Temperament
UnderLang	Understanding of language
Vocab	Vocabulary
Write	Writing

The computation method for PCA assumes that the level of measurement is continuous, allowing correlations to be computed. However, this requirement is often relaxed so that ordered data, most notably from Likert-type scales, can be used (see Hutcheson & Sofroniou, 1999, for a full discussion of this issue). It is also recommended that PCA be conducted on standardised scores, particularly when the variables are measured on different scales (e.g., when some variables are measured using 5-point scales, whilst others are measured as percentages). The following analysis uses standardised 5-point scores for computing the principle components (it is important to be aware of how the statistical software treats the data and whether scores are standardised by default).

A Correlation Matrix of the Data

The example at the beginning of this entry (Table 1) shows a data set where the underlying structure can be discerned directly from the correlation matrix. As a preliminary step in the data analysis process, it is often useful to inspect the correlations between the variables. Figure 1 shows a correlation matrix from a small selection of variables from the example data set.

Figure 1. Correlation matrix of selected variables in the example data.

Pearson correlations:

	Comp	Draw	MotSk	SocInt01	SocInt02	SentCom
Comp	1·0000	0·5361	0·4367	0·4756	0·4407	0·7110
Draw	0·5361	1·0000	0·5295	0·4157	0·3839	0·4454
MotSk	0·4367	0·5295	1·0000	0·2888	0·3068	0·3991
SocInt01	0·4756	0·4157	0·2888	1·0000	0·7632	0·4203
SocInt02	0·4407	0·3839	0·3068	0·7632	1·0000	0·3413
SentCom	0·7110	0·4454	0·3991	0·4203	0·3413	1·0000

Number of observations: 293

Pairwise two-sided p-values:

	Comp	Draw	MotSk	SocInt01	SocInt02	SentCom
Comp		<·0001	<·0001	<·0001	<·0001	<·0001
Draw	<·0001		<·0001	<·0001	<·0001	<·0001
MotSk	<·0001	<·0001		<·0001	<·0001	<·0001
SocInt01	<·0001	<·0001	<·0001		<·0001	<·0001
SocInt02	<·0001	<·0001	<·0001	<·0001		<·0001
SentCom	<·0001	<·0001	<·0001	<·0001	<·0001	

It is clear from the correlation matrix in Figure 1 that all variables are significantly correlated. This is to be expected with these data, as some children tend to get higher scores than others at all tasks, either due to being at different developmental stages or due to a difference in general ability or "IQ." From this output, it is not easy to discern patterns of correlation that may indicate more subtle underlying structure in the data. This is typical of many data sets where the underlying structure is difficult to discern directly from the correlation matrix. In this case, PCA is applied to provide a more detailed decomposition of the data.

Measure of Sampling Adequacy

PCA identifies the underlying structure in a data set even if there is little structure to be identified (e.g., if the marks the children achieved in the individual tests are not dependent on a smaller subset of core abilities). It is useful, therefore, to determine whether the variables in the sample are related in such a way that components underlying the structure can be identified. It is important to know the strength of the structure underlying the

relationships in the data (and thus, the strength of the possible underlying “causes” of the correlations) and thereby determine whether a PCA can be appropriately applied.

A useful method for determining the appropriateness of running a PCA is to compute a measure of sampling adequacy (MSA) which are commonly known as Kaiser-Meyer-Olkin, or KMO, statistics. Such measures were proposed by Henry F. Kaiser (1970) and are based on an index which compares correlation and partial correlation coefficients (the equations used to compute MSA statistics are provided in Hutcheson & Sofroniou, 1999). Measures of sampling adequacy may be computed for multiple variables and for individual variables. One can, therefore, investigate how “well” the overall data set can be represented in terms of an underlying structure and also how strongly individual variables contribute to this structure.

Similar to correlations, KMO statistics take values between 0 and 1. For large KMO values, there is likely to be a pattern of correlation indicating a strong underlying structure to the data, which implies that PCA might be an appropriate technique to use. For low KMO values (approaching 0), the relationships in the data are quite diffuse, making it unlikely that the variables will form distinct components suggesting PCA is not an appropriate technique to apply. Kaiser suggests a descriptive method of interpreting the KMO statistics (Kaiser, 1974) and this is shown in Table 3.

Table 3. Interpretation of the Kaiser-Meyer-Olkin (KMO) statistics.

KMO statistic	Interpretation
In the .90’s	Marvellous
In the .80’s	Meritorious
In the .70’s	Middling
In the .60’s	Mediocre
In the .50’s	Miserable
Below .50	Unacceptable

Measures of sampling adequacy are commonly reported as part of the standard output of PCA. In R, the library psych (Revelle, 2018) provides a function to compute overall and individual MSA scores. Figure 2 shows the MSA statistics for the PCA example data set computed using the psych library.

Figure 2. MSA statistics (19 variables).

```
Kaiser-Meyer-Olkin factor adequacy
Overall MSA =  0·91
MSA for each item =
   Active     Artic     Atten      Comp     Coord       Day      Draw
     0·89      0·92      0·85      0·95      0·90      0·43      0·92

  LangExp LangUnder      Math     Month     MotSk    NewSit   SentCom
     0·95      0·94      0·93      0·29      0·94      0·95      0·92

  SocInt01  SocInt02      Temp     Vocab     Write
     0·88      0·92      0·73      0·93      0·92
```

The overall MSA score of 0.91 indicates a “marvellous” result, suggesting a strong underlying structure to the data set. It is useful, however, to also look at the MSA statistics for individual variables. The MSA statistics for each item show that most of them are above 0.7 and look to be part of a good solution. The variables *Day* and *Month*, however, have relatively low individual MSA scores (0.43 and 0.29, respectively, which are deemed to be unacceptable using Kaiser’s terminology) which indicates that these variables are not strongly related to other variables in the data set and might be superfluous for defining any underlying structure. It is, therefore, worth looking at these variables in more detail.

The first thing to note about the variables *Day* and *Month*, is that they are in fact unordered categorical data and should not, therefore, even have been included in this analysis as any correlations are, essentially, meaningless (although unordered categorical variables should not be part of a PCA, it is relatively common mistake for them to be included in analyses, particularly when data have been incorrectly coded as numbers). These variables are also not likely to be of interest in regards to the abilities of the children, as whether a child was born on a Monday or a Thursday, or which month they were born in, is not of theoretical interest as they are unlikely to be meaningfully related to any other variables in the data or the proposed underlying structure. These two variables have been wrongly included in the analysis and should, therefore, be removed. Re-running the MSA statistics on the 17 variables that are left does not significantly change the overall MSA score which still indicates a “marvellous” solution (a KMO score of 0.92) but now indicates that all variables have individual MSA scores above 0.7.

Figure 3. MSA statistics (17 variables).

```
Kaiser-Meyer-Olkin factor adequacy
Overall MSA =  0.92
MSA for each item =
   Active      Artic      Atten       Comp      Coord       Draw
     0.89       0.93       0.85       0.95       0.90       0.93
  LangExp  LangUnder       Math      MotSk     NewSit    SentCom
     0.95       0.94       0.93       0.94       0.97       0.93
 SocInt01   SocInt02       Temp      Vocab      Write
     0.88       0.92       0.73       0.93       0.92
```

There are no hard-and-fast rules about when to remove variables from an analysis based purely on the KMO scores. The inclusion of variables should also be decided on grounds of analytical appropriateness and be at the discretion of the analyst. The MSA scores are, however, useful to check for inappropriate variables, those which have been miscoded and items which may be of theoretical importance, but lack enough items to properly define (an item may have a low MSA score if it is the only item measuring a particular underlying construct; if this construct is important it may be retained or further items that are related to it may be included in the analysis).

Running the PCA

Once the variables that are to be used in the analysis have been carefully selected (based on theoretical considerations and the KMO statistics), the individual components that may define the underlying structure in the data can be determined using PCA. PCA identifies linear combinations of the observed variables with the first principal component (PC01) being the linear combination that accounts for the largest amount of variance in the sample. The second principal component (PC02) is the linear combination which accounts for the maximum amount of the remaining variation. Successive components explain progressively smaller portions of the total sample variance, until all variance has been accounted for (when the number of components is the same as the number of variables). All components are uncorrelated with each other (orthogonal). Essentially, a PCA transforms a set of *n* correlated *variables* into a set of *n* uncorrelated *components*.

Figure 4 shows the component loadings from a PCA where the 17 correlated variables in the example data set have been transformed into 17 uncorrelated components. The component loadings show the relationship between each of the variables and each of the components. It should be noted that the signs (positive or negative) of the loadings are more or less arbitrary and do not have any substantive meaning. The signs within a component are, however, important as they indicate comparative differences between the individual variables and each component. The R program “princomps” from the stats library was used to compute the component loadings as shown in Figure 4 (R Core Team, 2018).

Figure 4. Component loadings.

Component loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Active	-0.230	-0.270	0.265	0.351	-0.080	0.037	-0.146	-0.183	0.160
Artic	-0.253	0.357	0.148	-0.097	-0.073	-0.072	-0.063	-0.254	0.196
Atten	-0.209	-0.248	0.224	-0.474	0.208	0.191	0.084	-0.063	0.630
Comp	-0.287	0.174	-0.044	0.043	0.149	0.022	0.067	0.543	-0.150
Coord	-0.255	-0.155	-0.369	-0.027	-0.158	0.215	0.036	0.027	0.076
Draw	-0.244	-0.152	-0.368	-0.030	-0.204	0.266	-0.191	-0.030	-0.100
LangExp	-0.273	0.276	0.110	-0.085	0.126	0.045	-0.159	-0.155	-0.328
LangUnder	-0.280	0.215	0.051	0.028	0.056	0.137	0.160	0.569	0.206
Math	-0.187	-0.053	-0.314	0.244	0.796	-0.187	0.205	-0.236	0.011
MotSk	-0.199	-0.079	-0.330	-0.190	-0.244	-0.809	-0.094	0.087	0.190
NewSit	-0.253	-0.055	0.090	0.124	-0.334	-0.037	0.840	-0.200	-0.142
SentCom	-0.276	0.317	0.061	-0.102	-0.002	-0.048	-0.138	-0.247	-0.090
SocInt01	-0.242	-0.244	0.265	0.377	-0.040	-0.100	-0.200	-0.033	0.033
SocInt02	-0.231	-0.310	0.260	0.204	0.022	-0.137	-0.168	0.248	-0.145
Temp	-0.149	-0.371	0.217	-0.571	0.136	-0.069	0.034	0.015	-0.511
Vocab	-0.274	0.330	0.108	-0.018	-0.089	0.017	-0.115	-0.101	-0.003
Write	-0.240	-0.157	-0.393	0.004	-0.067	0.304	-0.132	-0.130	-0.058

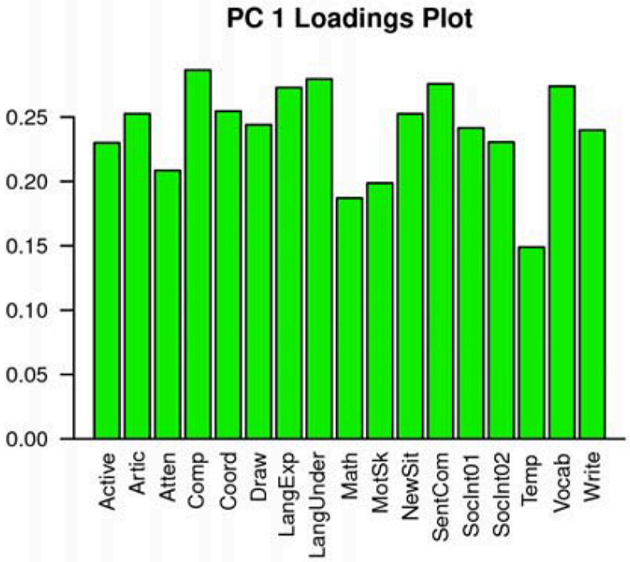
	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17
Active	0.514	0.218	-0.031	-0.226	-0.080	0.184	0.427	0.082
Artic	-0.065	0.104	-0.475	-0.083	-0.369	0.325	-0.356	-0.212
Atten	-0.061	-0.058	0.055	0.044	-0.043	-0.338	0.022	0.071
Comp	0.276	-0.022	0.016	0.246	-0.580	-0.237	0.094	-0.022
Coord	-0.413	-0.081	0.396	-0.228	-0.310	0.378	0.219	-0.152
Draw	0.226	-0.559	-0.389	-0.204	0.123	-0.189	-0.106	0.055
LangExp	-0.089	0.246	0.247	-0.583	0.024	-0.421	-0.080	0.046
LangUnder	0.082	0.094	0.011	-0.203	0.501	0.330	-0.173	0.111
Math	-0.002	-0.102	-0.090	-0.019	0.082	0.081	0.023	-0.075
MotSk	0.095	0.102	0.061	-0.068	0.071	-0.100	0.018	0.022
NewSit	-0.022	-0.007	-0.041	0.026	0.027	-0.156	-0.032	0.053
SentCom	-0.065	-0.221	0.154	0.326	0.026	0.205	0.141	0.688
SocInt01	0.006	-0.244	0.420	0.179	-0.002	-0.023	-0.574	-0.139
SocInt02	-0.605	0.060	-0.427	0.033	0.045	-0.101	0.182	0.118
Temp	0.186	-0.006	0.010	0.041	0.053	0.340	-0.046	-0.152
Vocab	-0.029	-0.140	0.056	0.304	0.347	-0.102	0.405	-0.602
Write	0.024	0.631	-0.052	0.408	0.134	-0.080	-0.192	0.033

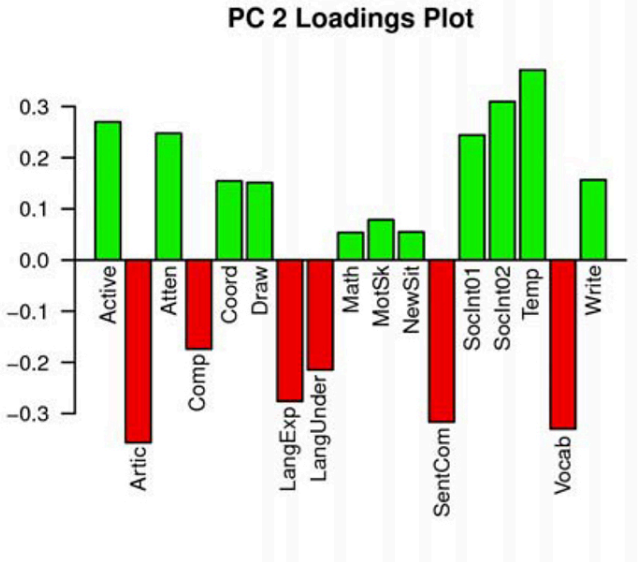
This is a standard output for a PCA and shows the factor loadings for each variable on each component. In this analysis, as there are 17 variables represented by 17 components, all of the variation in each variable is accounted for (we have, therefore, not “lost” any information by transforming the variables into components) and the squared loadings for any particular variable will sum to 1.0. For example, for the variable *Active*:

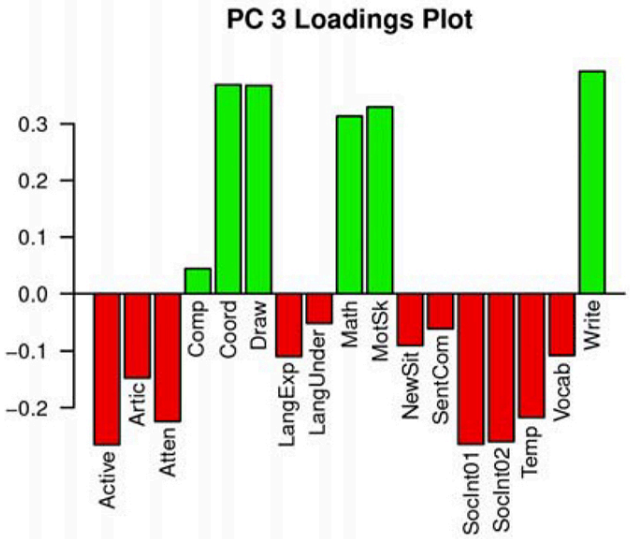
$$-0.230^2 + -0.270^2 + 0.265^2 + 0.351^2 + -0.080^2 + 0.037^2 + -0.146^2 + -0.183^2 + 0.160^2 +$$
$$0.514^2 + 0.218^2 + \text{-}0.031^2 + -0.226^2 + -0.080^2 + 0.184^2 + 0.427^2 + 0.082^2 = 1.0$$

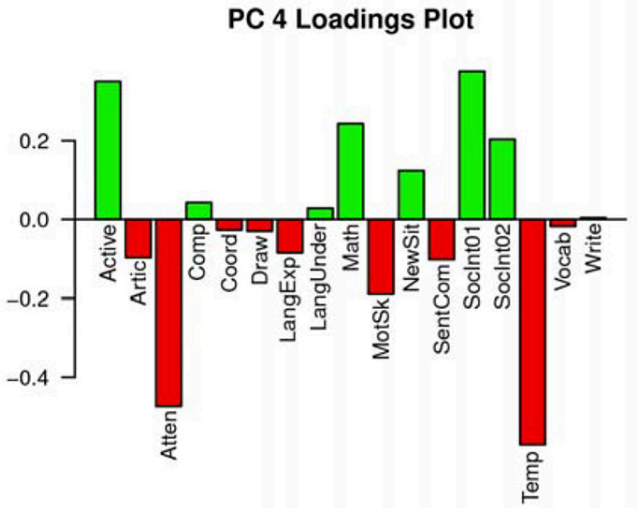
An overall impression of the relationship between the variables and the components is difficult to discern directly from the tabular output and many software packages provide graphical tools to make this task easier. Figure 5 shows the component loading table in graphical form drawn using the prcomp command from the R stats library (note that the signs do not always match between the tabular results and the bar chart; the values of the loadings and the comparative signs are, however, preserved.).

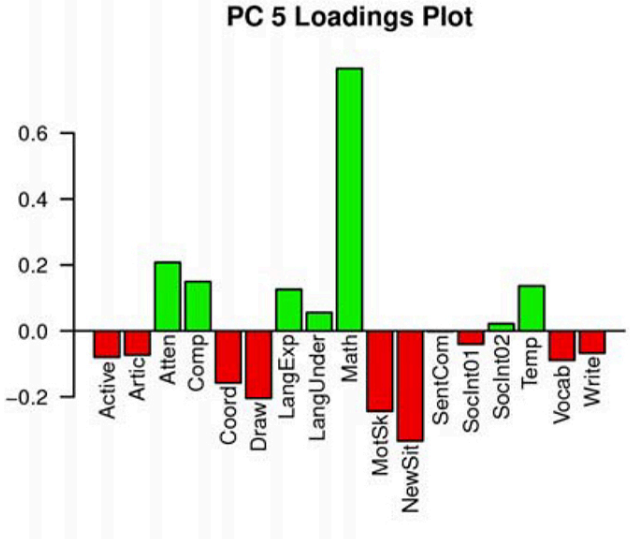
Figure 5. PCA component loadings.

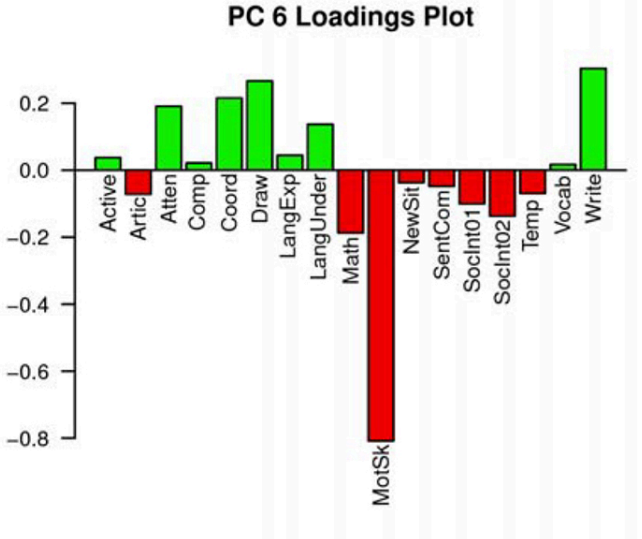












The bar charts of the loadings provides a lot of information in a relatively easy-to-digest format. All variables load in the same direction on principal component (PC) 1, which gives some evidence for a general relationship between children on all abilities (some children are much better or worse on all tasks than others). PC2 appears to distinguish between the more social skills (social interaction, activity, temperament) and linguistic skills (articulation, language, and vocabulary). PC3 distinguishes between social and visual skills (coordination, drawing, and writing), whilst PC4 provides some distinction between social interaction and activity scores with attention and temperament. It is interesting to note that PC5 and PC6 distinguish single variables (mathematics and motor skills) making these components not really indicative of a general underlying structure in the data. A quick glance at these graphs suggest not only the definitions of the components, but also that four components may be useful in identifying the underlying structure.

At this stage of the analysis, the 17 correlated variables (*Active* to *Write*) have been transformed into 17 uncorrelated components (Comp.1 to Comp.17). This is demonstrated in Figure 6, which shows the correlations between the first four variables and the first four components. The variables are correlated with each other, whereas the components are all uncorrelated (orthogonal).

Figure 6. Correlation matrix of variables and components.

Variables:

	Active	Artic	Atten	Comp
Active	1.00	0.35	0.47	0.43
Artic	0.35	1.00	0.36	0.65
Atten	0.47	0.36	1.00	0.37
Comp	0.43	0.65	0.37	1.00

Components:

	PC1	PC2	PC3	PC4
PC1	1.00	0	0	0
PC2	0	1.00	0	0
PC3	0	0	1.00	0
PC4	0	0	0	1.00

Contribution of the Components

Information about the components is provided in Figure 7, which shows the amount of variance in all the data that is accounted for by each component. As there are 17 components, the total variance sums to 17 (8.362 + 2.013 + … + 0.124 = 17.0). The components account for different amounts of variance in the data with the first component accounting for nearly 50% of the total variance (8.362 out of 17), whilst the second accounts for about 12% (2.013 out of 17). The remaining components account for successively smaller amounts with the final component accounting for only 0.07% (0.124 out of 17).

The lower panel in Figure 7 shows the importance of the components, at least with regards to the amount of deviance they account for. The lower panel also shows the individual and cumulative proportion of variance accounted for by each component. For example, the first two components account for about 60% of the variation in the data (0.492 + 0.118 = 0.610, this number can also be seen in the third row), whereas the first four components account for about 76% (0.492 + 0.118 + 0.093 + 0.059 = 0.762).

Figure 7. Component variances.

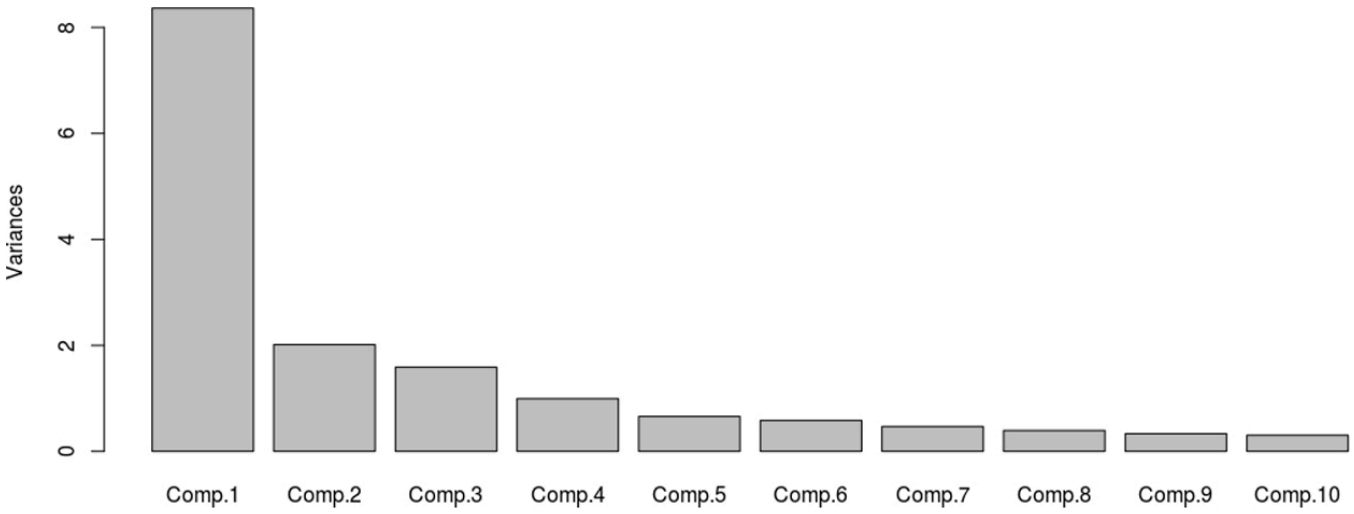
```
Component variances:
 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8
Comp.9
  8.362   2.013   1.586   0.996   0.658   0.582   0.469   0.393
0.333
Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
  0.303   0.274   0.204   0.195   0.182   0.169   0.158   0.124

Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Comp.7
Standard deviation      2.892  1.419  1.259  0.998  0.811  0.763
0.685
Proportion of Variance  0.492  0.118  0.093  0.059  0.039  0.034
0.028
Cumulative Proportion   0.492  0.610  0.704  0.762  0.801  0.835
0.863
                       Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
Standard deviation      0.627  0.577   0.550   0.523   0.452   0.442
Proportion of Variance  0.023  0.020   0.018   0.016   0.012   0.012
Cumulative Proportion   0.886  0.905   0.923   0.939   0.951   0.963
                       Comp.14 Comp.15 Comp.16 Comp.17
Standard deviation       0.427   0.411   0.398   0.352
Proportion of Variance   0.011   0.010   0.009   0.007
Cumulative Proportion    0.974   0.983   0.993   1.000
```

Selecting the Number of Components

From Figure 7, one can see that most of the variation in the data can be described using a limited subset of components. For example, if just the first three components were included in an analysis, over 70% of the variance would be accounted for. An acceptable model of the data may be able to be made using relatively few components. The question is, How many components are needed to adequately describe the data?

This question, as it turns out, is not an easy one to answer, as it depends on how interpretable the components are, the aims and objectives of the analysis, the size of the sample, and the number of variables considered. Numerous methods have been proposed in the literature for selecting the "optimum" number of components needed to describe a data set (see Jolliffe, 2002, for a comprehensive review of these). These include selecting those components that collectively account for a certain percentage of the variation in the whole data set (say, 80%), selecting just those components that individually account for a certain amount of the variance (usually a component standard deviation above 1.0, as this is the amount of deviation associated with a single variable; because this is a very easy statistic on which to base the decision, it is quite often the default used in statistical packages), by investigating the pattern in the variance accounted for by each component using a scree plot, or by carefully interpreting the components and selecting just those that are meaningful with respect to the objectives of the analysis. A scree plot is, perhaps, the most widely used technique for selecting the number of components and is commonly provided as a default graphic. A scree plot for the example data is shown in Figure 8.

Figure 8. A scree plot.



The scree plot shows the variances for the first 10 components in descending order. Interpreting the plot is quite easy—one just looks for the point in the graphic where the gradient becomes constant. For this graphic, a relatively straight line can be constructed linking Components 5 to 10. Component 4, however, appears to account for a slightly greater proportion of the variance and indicates the point where the gradient changes. This might indicate that Component 4 is measuring something general that accounts for some additional

variance—maybe an underlying cause. The scree plot suggests that a four-component solution may be a good choice to model these data. Other methods suggest different numbers of components to retain. For example, using the criteria that an individual component accounts for more than an individual variable (a standard deviation of above 1) results in three components being retained whilst accounting for over 80% of the total variance results in five components being retained. It is interesting to note that the bar charts of the component loadings (see Figure 5) also suggest a four-component solution as additional components tend to distinguish single variables as opposed to underlying structures.

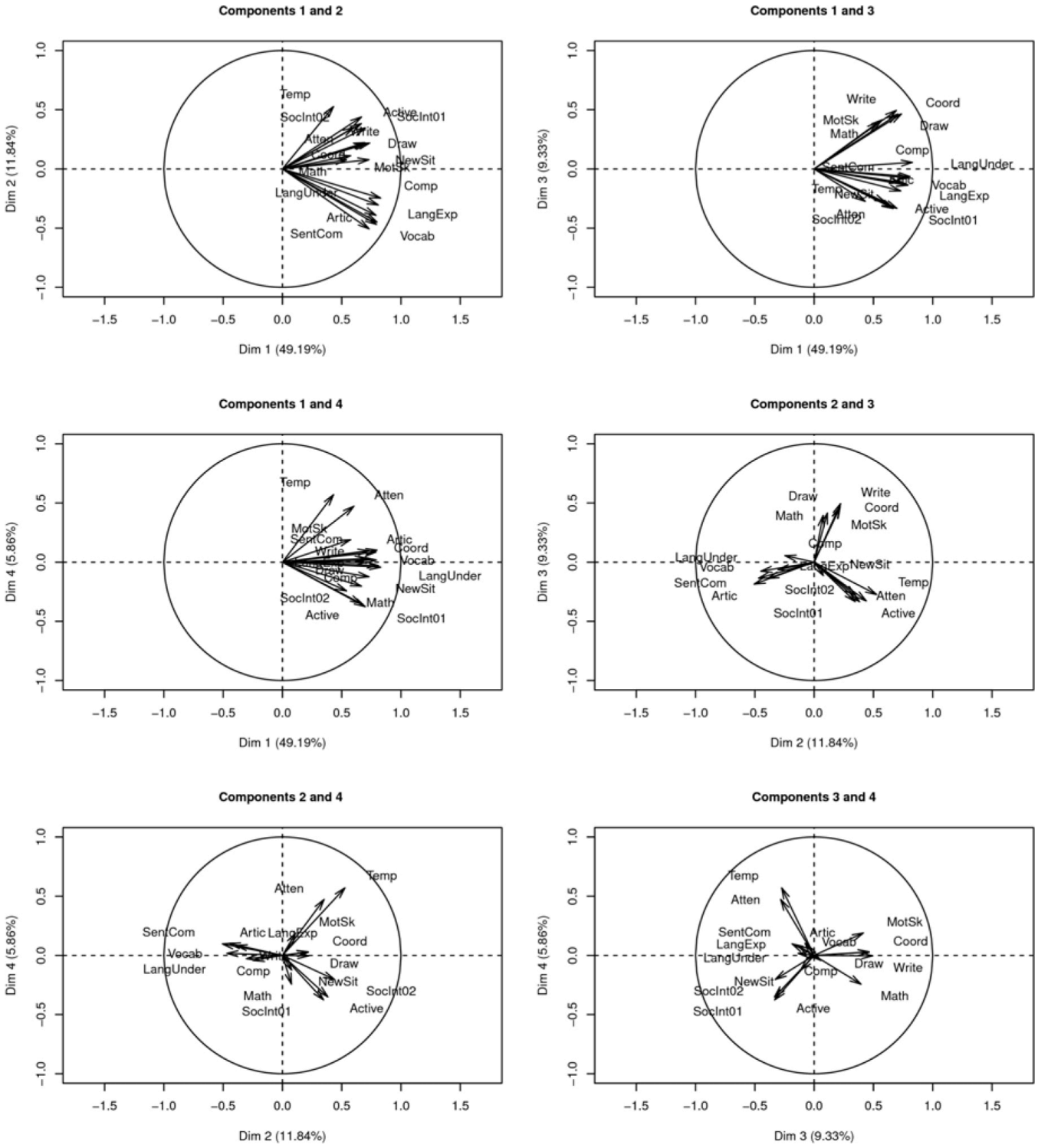
In practice, the selection of the number of components is an art form. There are no strictly correct answers. The number of components need to be decided on the basis of what make sense in the analysis. For this example, we will select four components to represent our data which account for 76% of the variance in the entire data set.

Interpreting the Components

At this stage in the analysis, although we have computed and selected four components to describe the data, there is little indication as to the underlying structure. The component loadings shown in Figures 4 and 5 provide some indication of which variables load highly on particular components, but it is not a simple matter to discern any underlying structure using just this information, as the structure is often realised through a combination of components. To discern the structure more clearly, it is useful to show the relationships between the components using graphical displays. There are many different displays available, but perhaps the most common are simple biplots, which show the loadings for two of the PCs and include labels for each of the original variables (see Gower & Hand, 1996). These plots have been produced here using the R FactoMinR library (Le et al., 2008) run through the Shiny web interface (Chang, Cheng, Allaire, Xie, & McPherson, 2017) which provides a simple method for producing these graphics. (There are many graphics which can be used to illustrate the results of PCA and a large number of choices within this package which can also be easily accessed using the FactoMineR plugin (Husson et al., 2016) for the R-commander interface (Fox & Bouchet-Valat, 2018). The graphics shown in Figure 9 have used the default graphical options. Analysts are encouraged to investigate which ones are best suited for their particular needs.)

Biplots show the component loadings for two components at a time (one on each axis) and indicate the loading for each variable using an arrow. The length of the arrow indicates the size of the loading and all variables are named on the graphic. At first glance, the biplots can appear to be very busy and difficult to read, but they are an incredibly rich source of information and deserve to be studied in some detail. All biplots for the four components retained in our analysis are shown in Figure 9.

Figure 9. Component biplots.



The biplots show each component (named as Dimensions, or Dim) and the proportion of variance associated with this component is shown in the axis label. The PCA aims to identify the underlying structure in the data and it is evidence for this structure that we look for within the plots. A good place to start is to look for variables that are in close proximity. For example, in the top right biplot which shows the loadings on Components 1 (*x*-axis) and 3 (*y*-axis), *Motor Skill*, *Writing*, *Coordination*, and *Drawing* are close together and form one cluster whilst *Comprehension*, *Language Understanding*, *Vocabulary*, and *Expressive Language* are also close together and form a different cluster. These clusters of variables are linked and may identify an

underlying relationship. The first cluster of variables may indicate a visual-spatial component (it is interesting to note that the variable ***Maths*** appears close to the visual-spatial cluster of variables which may indicate that mathematical ability is associated with visual-spatial skills), whilst the second may indicate a linguistic component. Some of the graphs show these clusters better than others; perhaps the best differentiation between the clusters can be seen in the middle-right graphic, which shows Components 2 and 3. A more detailed view of this biplot is shown in Figure 10, which shows the original biplot of Components 2 and 3 with three ellipses added to help identify the clusters. It is quite clear that there are three distinct groups that can be identified. The variables which cluster in these groups are shown in Table 4.

Figure 10. A biplot of Components 2 and 3.

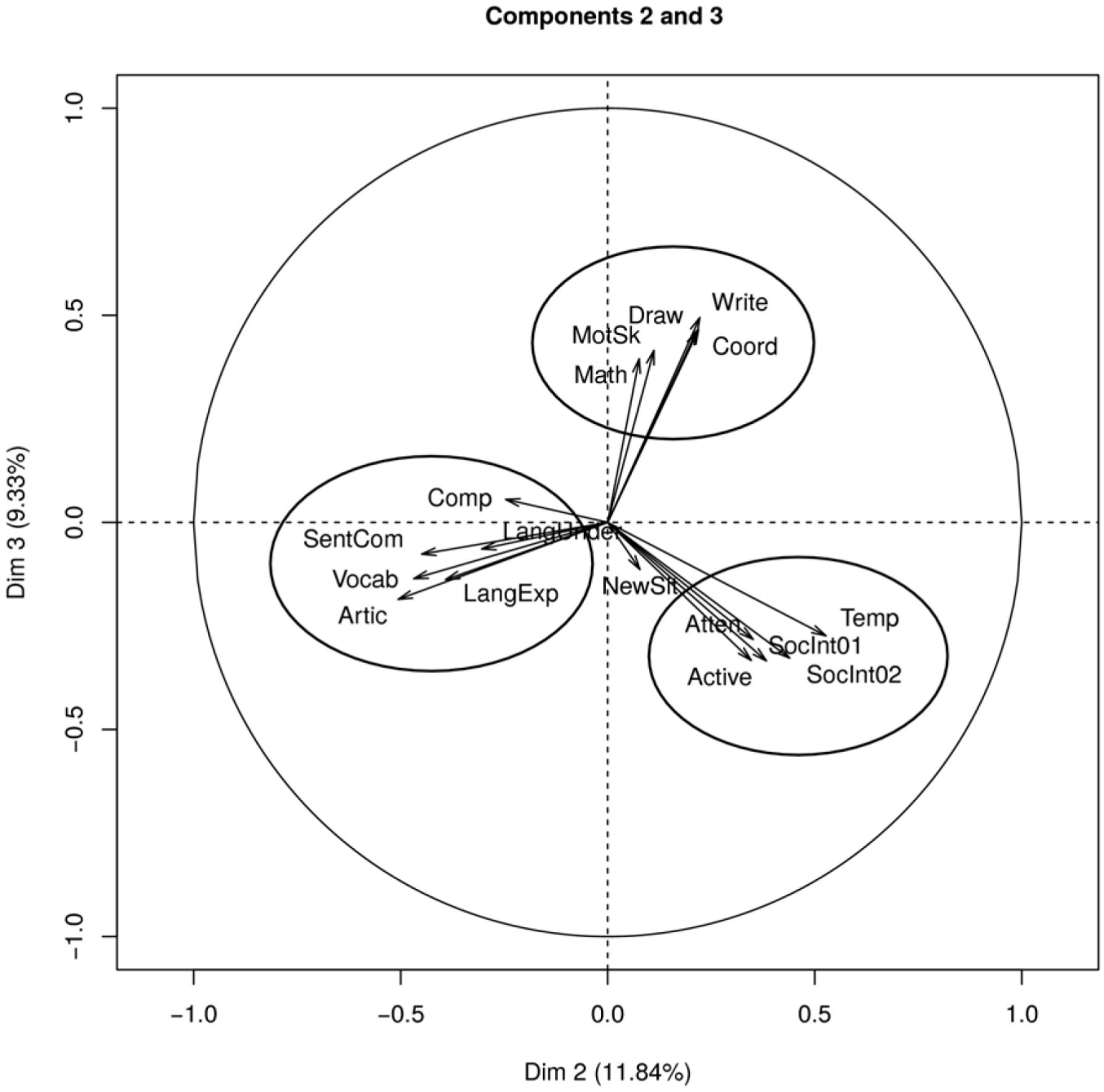


Table 4. Membership of the clusters.

Cluster A (Visual-Spatial)	Cluster B (Social and Behavioural)	Cluster C (Linguistic)
Writing	Social interaction	Articulation
Drawing	Temperament	Vocabulary
Coordination	Activity	Sentence completion
Motor skills	Attention	Language
Maths		Comprehension

The variables within each of the groups are closely related (based on substantive knowledge about the variables) and can be broadly identified as visual-spatial abilities (Group A), social and behavioural (Group B), and linguistic abilities (Group C). These groups may identify the underlying abilities that are responsible for the correlations in the data set. The abilities of children may, at least to some degree, be determined by these three core abilities that are relatively distinct (at least in the sense that they can be differentiated by the analysis). It might be useful to think of the abilities of children in relation to these three main underlying constructs, rather than the 17 individual variables that were originally assessed.

There are finer distinctions that can also be derived from the biplots. For example, the graphics in Figure 9 which include Component 4 show a noticeable difference between attention and temperament (which make up one cluster) and the other social interaction variables (which make up another cluster). This could indicate different skills associated with these aspects of the social component; a result which is also suggested in the PC4 loadings plot in Figure 5. It may be useful to further investigate the reasons for this clustering and if more clusters may usefully be extracted from the data.

In addition to identifying the main clusters in the data, the graphics are also useful in describing individual variables. For example, it is interesting to look at mathematics and the evidence that this is a distinct skill (at least with respect to the other skills measured in this study), or whether it is part of another, more general ability. Figure 10 suggests that it is most strongly related to the visual-spatial group of abilities, but there is also evidence that it is distinct from these, given that it appears quite separate from this cluster in the graphic that compares Components 3 and 4 in Figure 9. The component loadings for PC5 in Figure 5 also suggest that mathematics is distinct from the other variables.

The identification of the underlying structure in the data from the component loadings is not a simple task. What is clear is that the underlying structure is realised as a complex interaction of the component loadings and requires careful interpretation.

Factor Analysis

The underlying structure in the data can be identified by analysing the PCs using the tools outlined in this entry. The PCs, however, do not directly represent the underlying structure in the data as this is realised as a complex interaction of the component loadings. It is often useful to represent the underlying structure as variables (e.g., a factor to represent visual-spatial ability, one to represent social skills and another to represent linguistic abilities) so that these can be used in other analyses (in regression models, for example; see Jolliffe, 1982; Maindonald & Braun, 2003). This can be achieved by using the technique of factor analysis (Gorsuch, 1983; see Hutcheson & Sofroniou, 2008, for a simple introduction). Factor analysis maintains the relationship between the variables, but changes the axes (the components) using a process known as rotation. Factor analysis maximises the component loadings by rotating the axes whilst leaving the relationships between the variables unchanged. The rotation method may keep the resulting factors uncorrelated (orthogonal rotation) or may allow for some correlation between them (oblique rotation). Oblique rotations may be applied particularly when the resulting factors may be expected to correlate in the population. For example, in the present analysis, linguistic abilities and social skills may be related, which can be represented as a correlation between the factors. There are a number of different rotation methods that may be used, including ones that minimise the number of variables that load highly on individual factors in order to enhance interpretability to ones that minimise the number of factors (see Bernaards & Jennrich, 2006, for a discussion of available methods). A demonstration of using factor analysis on the same data set as used here can be seen in *The Multivariate Social Scientist: An Introduction to Generalized Linear Models* (1999) by Graeme D. Hutcheson and Nick Sofroniou, who also provide graphical illustrations of orthogonal and oblique rotations.

In addition to factor analysis, there are a number of other techniques that are similar to PCA and may also be used to classify data. These methods include cannonical correlation analysis for continuous data (Knapp, 1978), correspondence analysis which is traditionally applied to contingency table data (see Greenacre, 2007), and multiple correspondence analysis which provides the counterpart of PCA for categorical data (Greenacre & Blasius, 2006).

Conclusion

In the worked example included in this entry, the 17 variables which were originally selected to take part in the analysis could be reduced to just four components, which together accounted for over 75% of the variance in the data. In addition to this data reduction, the analysis also uncovered (or at least suggested the existence of) an underlying structure in the data set that is likely to be of theoretical interest with respect to the abilities of children. The analysis also raised some interesting questions for further research, particularly around the relationships between the components and individual variables (e.g., performance in mathematics). As seen from this entry, PCA can simplify data sets and reveal underlying structure and is a valuable technique for the

exploration and description of multivariate data.

References

Bernaards, C. A., & **Jennrich, R. I.** (2006). Gradient projection algorithms and software for arbitrary rotation criteria in Factor Analysis. *Educational and Psychological Measurement*, 65, 676–696.

Chang, W., **Cheng, J.**, **Allaire, J.**, **Xie, Y.**, & **McPherson, J.** (2017). Shiny: Web application framework for R. R package version 1.0.5. Retrieved from https://CRAN.R-project.org/package=shiny

Crawley, M. J. (2013). *The R book* (2nd ed.). Chichester, England: Wiley.

Fox, J., & **Bouchet-Valat, M.** (2018). Rcmdr: R Commander. *R package version*2.4–4.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). London, England: Lawrence Earlbaum Associates.

Gower, J. C., & **Hand, D. J.** (1996). *Biplots*. London, England: Chapman and Hall.

Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). London, England: Chapman & Hall/CRC.

Greenacre, M., & **Blasius, J.** (Eds.). (2006). *Multiple correspondence analysis and related methods*. London, England: Chapman & Hall/CRC.

Husson, F., **Josse, J.**, & **Le, S.** (2016). RcmdrPlugin.FactoMineR: Graphical user interface for FactoMineR. R package version 1.6-0. Retrieved from https://CRAN.R-project.org/package=RcmdrPlugin.FactoMineR

Hutcheson, G. D., & **Sofroniou, N.** (1999). *The multivariate social scientist: An introduction to generalized linear models*. London, England. SAGE.

Hutcheson, G. D., & **Sofroniou, N.** (2008). Factor analysis. In **L.Moutinho** & **G. D.Hutcheson** (Eds.), *The sage dictionary of quantitative management research* (pp. 117–121). London, UK: SAGE.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 31, 300–303.

Jolliffe, I. T. (2002). *Principal components analysis* (2nd ed.). Berlin, Germany. Springer.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, 35, 401–415.

Kaiser, H. F. (1974). An index of factor simplicity. *Psychometrika*, 39, 31–36.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85, 410–416.

Le, S., **Josse, J.**, & **Husson, F.** (2008). FactoMineR: An R package for multivariate analysis. *Journal of*

Statistical Software, 25, 1–18. doi:10.18637/jss.v025.i01

Maindonald, J., & **Braun, J.** (2003). Data analysis and graphics using R; An example-based approach. *Cambridge series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych Version = 1.8.3.