Semi-automated Quality Assurance for Domain-Expert-Driven Data Exploration – An Application to Principal Component Analysis

Sandra Wartner, Manuela Wiesinger-Widi, Dominic Girardi, Dieter Furthner, Klaus Schmitt

Semi-automated Quality Assurance for Domain-Expert-driven Data Exploration – An Application to Principal Component Analysis[⋆]

Sandra Wartner[1], Manuela Wiesinger-Widi[1], Dominic Girardi[1], Dieter Furthner[2], and Klaus Schmitt[3]

[1] RISC Software GmbH, Research Unit Medical Informatics, Hagenberg, Austria
firstname.lastname@risc-software.at
[2] Department of Pediatrics and Adolescent Medicine, Salzkammergut-Klinikum, Vöcklabruck, Austria
[3] Department of Pediatrics and Adolescent Medicine, Kepler University Hospital, Linz, Austria

Abstract. Processing and exploring large quantities of electronic data is often a particularly interesting but yet challenging task. Both the lack of statistical and mathematical skills and the missing know-how of handling masses of (health) data constitute high barriers for profound data exploration – especially when performed by domain experts. This paper presents guided visual pattern discovery, by taking the well-established data mining method Principal Component Analysis as an example. Without guidance, the user has to be conscious about the reliability of computed results at any point during the analysis (GIGO-principle). In the course of the integration of principal component analysis into an ontology-guided research infrastructure, we include a guidance system supporting the user through the separate analysis steps and we introduce a quality measure, which is essential for profound research results.

Keywords: Data Quality · Guidance · Principal Component Analysis · Visual Analytics · Data Mining · Doctor-in-the-Loop.

1 Introduction

Due to the steadily rising amount of data in varying research domains, visual data analytics is becoming increasingly important. In complex research domains (such as biomedical research), deep integration of the domain expert into the data analysis process is required [1]. A major technical obstacle for these researchers lies not only in handling, processing and analyzing complex research data [2], but also in giving chapter and verse for exploiting results. Since conclusions can

[⋆] This work was supported in part by the Austrian Research Promotion Agency (FFG project number 851460) and by funds from the Strategic Economic and Research Program "Innovatives OÖ 2020" of the State of Upper Austria.

be no better than the received input, users also have to be aware of this GIGO ("Garbage-In-Garbage-Out") principle when applying analytics methods.

With our work we aim to assist the domain expert in the whole process of data modeling, processing, analysis, and interpretation. In particular we are aspiring to increase interpretability and understanding of advanced analysis techniques, such as the Principal Component Analysis (PCA), which serves as an example for evaluating our approach. Preliminary work on this topic by Wartner et al. [3] has focused on the integration of basic PCA functionality, enabling its use by domain experts without assistance of a data scientist. Moreover, it has introduced the concept of quality of the result, by a preliminary selection of certain quality criteria, such as the sample size, the ratio between the number of observations and variables, and the properties of the correlation matrix. In this paper, an ascertainment of the quality criteria summarized and combined in an assessment scheme is presented. Finally, the guided PCA is applied on data from the MICA (Measurements for Infants, Children, and Adolescents) project which was performed in cooperation with the *Kepler University Hospital Linz.*

1.1 An Ontology-based Research Platform

In order to address the issue of handling, processing, and analyzing complex research data, we have been working on an ontology-based research platform for domain-expert-driven data exploration and research. The key idea behind the platform is that, while being a completely generic system, it can be adapted to any specific research domain by modeling its relevant aspects (classes, attributes, relations, semantic rules, constraints, etc.) in the form of a domain ontology. The whole system adapts itself to this domain ontology at run-time and appears to the user like an individually developed system. Moreover, the elaborate structural meta-information about the research data is used to actively support the domain-expert in challenging tasks such as data integration, data processing, and finally data exploration.

For a more detailed description on the platform itself and the usage of the domain-ontology for data exploration the reader is kindly referred to [3,4,5].

1.2 The Nuts and Bolts of Principal Component Analysis

Multidimensional data can be hard to explore and visualize. Methods such as the Principal Component Analysis (PCA) are used for simplifying this challenging task by decreasing the dimensions of the data set. Dimensionality reduction aims to reduce the number of variables in the data set without significant loss of information. The new (fewer) variables – in the case of PCA called principal components – are linear combinations of the original variables, capturing most of the variation of the original data set. For an introduction to PCA see for example [6,7,8].

Based on the centered covariance or correlation matrix of numeric variables, PCA works as a solution to the eigenvalue problem. The newly obtained orthogonal (i.e., uncorrelated) axes constitute linear combinations of the original variables and are referred to as principal components. The corresponding eigenvalue is a measure of importance of the principal component, describing the amount of variation in the original data explained by the principal component. When projecting data into a lower dimensional space, the primary motivation is to preserve most of the variation – the direction with maximum variation is found in the first principal component, while successive principal components account for less variance than the previous ones. As a result the least important principal components are discarded. What we receive is a data set with a reduced number of variables. The proportion of a component's variation, respectively, sums up to the explained variance of the new system. Eventually, the procedure is completed as soon as a predefined percentage of the original system's variance or number of components has been reached.

There are different ways of visualizing PCA results. Score plots depict the transformed data points, and thus are used to find patterns and clusters in the data and to detect outliers within the model, i.e., observations which scatter far from the data center (see also Figure 10). Ideally, the majority of data lie around the origin of the new coordinate system and spread just slightly. Further advantages arise when chronological sequences are to be analyzed. The analyst is able to determine when a process is getting out of control by tracking the time-related course of the transformed observations. In the event that the position of the observation is increasingly migrating from the origin and out of the control ellipse (see also Section 2.2), there is a high chance that one or more variables are taking on unfavorable values. Likewise, in non-sequential data (as in our experimental data), points of large deviation to the center of the hyperplane of the PCA model (and/or outside the control ellipse) differ more strongly from the rest of the observations. While score plots are used for examining the similar behavior of observations, loadings plots are used to investigate the influence of variables on a certain principal component (see also Figure 9). A variable's position (weight) close to zero indicates little importance, whereas high weights emphasize the contribution to the component. For pattern recognition, the relative positions of the variables to each other also play a pivotal role (close locations imply high correlation and vice versa). Merging both, the sample scores and variable loadings, in one visualization, the resulting visualization is called a biplot. In order to find observations which are not explained well by the model, samples can be colored by the squared prediction error (SPE) to graphically represent the size of the residuals.

2 Quality Measure

To the author's knowledge, no previous work has proposed the assessment of PCA result quality by a single metric. As a result, we develop an expansive evaluation scheme, modifiable for versatile types of analysis. In fact, we differentiate between the quality of a criterion and that of the entire data set, where the latter comprises the assessment of all quality criteria. Furthermore, the procedure of quality assessment also comprises quality criteria not assigned a grade, such as linearity or normality. However, the system makes domain experts aware of checking these requirements.

2.1 A Polymorphic Evaluation Scheme

Firstly, each quality criterion is graded as *good*, *ok*, and *bad*. Secondly, the overall quality is determined according to the subsequent rules:

- If more than 50% of the quality criteria are graded as *good*, and no *bad* has been given, the result quality is graded as *good*.
- If at least 50% of the quality criteria are graded as *ok* and no *bad* has been given, *ok* is awarded.
- If at least one criterion is graded as *bad*, the quality of the result is marked as inadequate (*bad* quality). This is because when at least one quality test fails, severe side effects might arise in interpreting the results.

It is worth noting that the choice of 50% is not restrictive, but merely reflects the experience of the authors in certain studies. One major benefit of this evaluation scheme is its polymorphism, as each data mining method and statistical analysis has most diverse requirements to the underlying data structure and quality. A tailor-made solution can be achieved by cobbling together various key figures in order to meet method dependent requirements. In the following section, the quality criteria specifically used for determining the PCA result quality are described.

2.2 Quality Criteria and User-interpretable Assessment Methods for Model Quality

This section provides an overview of the quality criteria specifically when applying PCA.

Sample size and ratio between number of samples and variables. A study by Osborne et al. [9] has shown that PCA results of data with both large sample sizes and high ratios had superior quality to those where the data had large sample sizes only. In the following table, as well as in the upper part of Figure 7, we adapt their proposed sample guidelines to our assessment scheme, where n denotes the number of observations and r the ratio between the number of observations and variables.

Table 1. Assessment scheme of the *sample size* quality criterion, where n denotes the number of observations. This classification is partially summarized from recommendations by [9].

Assessment	Sample size	Ratio
good	$n \geq 500$	$r \geq 10$
ok	$200 \leq n < 500$	$5 \leq r < 10$
bad	$n < 200$	$r < 5$

Communalities The communality r_x^2 of a variable x is computed as the sum of the squared correlations between the extracted principal components and the variable x. When using standardized data, the communality of a variable is computed as the sum of the squared *loadings* for this variable. It indicates how good the variable is explained by the extracted components. The closer the value to 1 is, the better the observed data for the variable is reflected in the model [10]. To the user, the communality for each selected variable is shown within the numeric output section (see last column in the result view section of Figure 8).

Correlation In a study conducted by Dziuban et al. [11], the importance of prior inspection of the correlation matrix, as well as the major interpretation pitfalls on examples of random data have been shown. It has been suggested that enough entries beyond the diagonal have to be greater or equal to 0.3 – raising a difficult question here in defining what is meant by “enough” [12]. To that end, the Kaiser-Meyer-Olkin test (KMO) has been introduced [13] to measure the sampling adequacy which gives, prior to analysis, an indication to the meaningfulness of applying PCA to the data set. It was later modified in Kaiser et al. [14] to the forms in Equations (1) and (2) to improve stability. A major advantage over other methods (e.g., the Bartlett’s test) is the possibility of simultaneously interpreting individual features and the overall quality of the correlation matrix [11].

Let (r_{ij}) denote the correlation matrix, (s_{ij}) the inverse of the correlation matrix and (q_{ij}) the *partial correlation* (or anti-image correlation) matrix of the input variables, where

$$q_{ij} := -\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

Then the overall measure of sampling adequacy (KMO) is defined as:

$$\mathrm{KMO} := \frac{\sum\limits_{i}\sum\limits_{j\neq i} r_{ij}^2}{\sum\limits_{i}\sum\limits_{j\neq i} r_{ij}^2 + \sum\limits_{i}\sum\limits_{j\neq i} q_{ij}^2} \tag{1}$$

The partial correlation estimates the relationship between two statistical variables while controlling for the effect of one or more other variables. Equation (2) provides the computation of the KMO for each variable separately:

$$\mathrm{KMO}(i) := \frac{\sum\limits_{j\neq i} r_{ij}^2}{\sum\limits_{j\neq i} r_{ij}^2 + \sum\limits_{j\neq i} q_{ij}^2} \tag{2}$$

In contrast to traditional measures of sampling adequacy, which have been proven to be unstable for poor data, this instability has been corrected through normalization by Olkin, hence any result value must lie between zero and one. Following Kaiser [14], the assessment scheme is shown in the correlation part in Figure 7. The closer the value is to 1, the more suitable the variables are.

Table 2. Rule for the interpretation of the overall KMO measure after Kaiser [14] and the related assessments integrated in the quality measure.

Assessment	Evaluation	KMO
good	marvelous	≥ 0.9
	meritorious	≥ 0.8
ok	middling	≥ 0.7
	mediocre	≥ 0.6
bad	miserable	≥ 0.5
	unacceptable	< 0.5

Linearity Another procedure of human supported quality assessment involves the inspection of the structure the data follows [15]. Linearity is given in case there is constant spread in data, i.e., data is homoscedastic, and no (strong) outliers are detected in the data set. This quality criterion is not assigned any grade, but is rather to be examined by the user. A simple and effective way of investigation is offered by a scatter plot matrix, showing $n \cdot (n-1)/2$ plots, where n is the number of variables. The variables names are written in a diagonal line from top left to bottom right. Each bivariate plot delineates the association between two variables of the data set – desirably the vast majority should show linear relationship, i.e., the plot should look like a line. Non-linear patterns between variables can not be detected by PCA (see Figure 1 for an example to interpret).

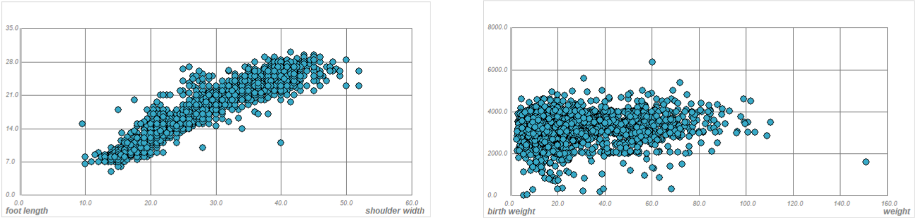


Fig. 1. Scatter plots showing the relationship between a pair of variables, each. In the first plot it is probably safe to say that there is a correlation between the variables, whereas the second plot does not show recognizable correlation.

Normally distributed data Although PCA can be performed on data that is not normally distributed, it might overlook patterns since it only handles first and second order dependencies like mean and variance. If the data is not normally distributed, higher order dependencies might be present but are not detected by PCA. Furthermore, independence of the components is only guaranteed in case of normality. If the data is not normally distributed, other methods like independent component analysis (ICA) may give further insight [16]. Automatic grading of normality is not included in the combined quality metric, but it is recommended to the user to investigate normality plots or normality tests (see also Thode [17]) in the detailed quality description listing. To easily assess whether data fits a normal distribution or not (and to which extent), an additional scatter-chart related visualization is provided in the research infrastructure rather than normality tests. In this plot, the normal theoretical quantiles are plotted against each variable (see Figure 2). According to the resulting curve characteristics, conclusions of the distribution can be drawn, including also supplementary information on kurtosis or skewness. If data follows a straight, linear pattern, it can be assumed that data is approximately normally distributed.

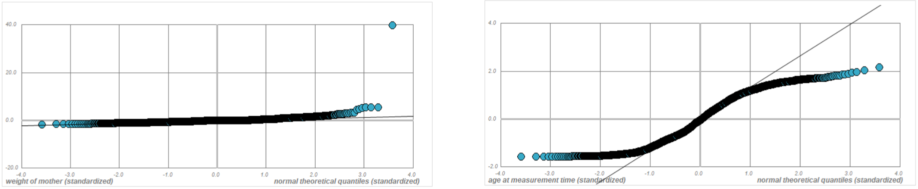


Fig. 2. The normality plots for the weight of the mother (on the left) and the age at measurement time (on the right). On the left-hand side graph, data follows approximately the normal distribution, except for the outlier on the top right corner. The right-hand side graph shows heavy tails, thus this variable doesn't follow the normal distribution.

Outliers For determining the suitability of samples for inclusion, outliers can be distinguished as:

- *outliers within the model*, detected by the control ellipse by the visual assessment of unusual points (see Figure 10), and,
- *outliers between the model and the measured data*, examined by the squared prediction error (SPE, see also Figure 3).

Like all linear methods PCA is very sensitive to outliers. Hence, a lot of attention has to be given to proper outlier handling.

Control ellipse The control ellipse is an addition to the score plot. Its aim is to visually identify potential outliers within the model. The control ellipse is derived via Hotelling's T^2, which indicates if a certain observation conforms to the mean of observations. In other words, it measures if an observation is in control. According to [6],

$$T^2 := zL^{-1}z = \frac{k(n-1)}{n-k} F_{k,n-k,\alpha},$$

where z is the score vector of said observation, L is the diagonal matrix of eigenvalues, k the number of principal components, n the number of observations,

α the significance level and $F_{k,n-k,\alpha}$ the critical value of the F-distribution with respect to the parameters k and $n-k$. Since we only draw two-dimensional control ellipses for principal components PC_i and PC_j, this formula reduces to

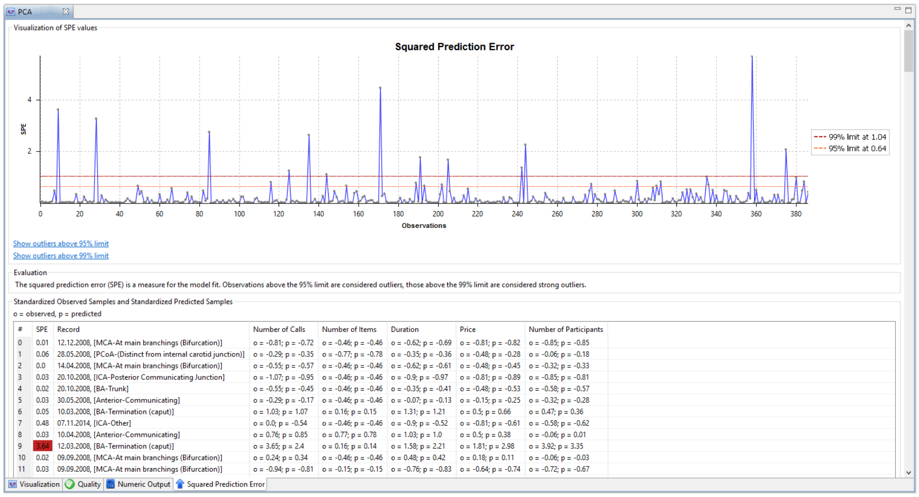
$$T^2_{i,j} = \frac{z_i^2}{\lambda_i} + \frac{z_j^2}{\lambda_j} = \frac{k(n-1)}{n-k}\, F_{k,n-k,\alpha},$$

with λ denoting the eigenvalues. Hence the control ellipse's center is 0 and its half-axes lengths $\mathrm{HA}(i)$ and $\mathrm{HA}(j)$ are

$$\mathrm{HA}(l) = \sqrt{\lambda_l\, F_{k,n-k,\alpha}\, \frac{k(n-1)}{n-k}}$$

for $l = i, j$. Most commonly used values for α are 0.05 or 0.01, depicting a control ellipse of 95% or 99%, respectively.

Squared prediction error The squared prediction error (*SPE* or *Q-statistic*) for an observation is defined as the squared difference between its actual value and the value predicted by the model, i.e., the scalar product of the error vector of the predicted vs. the actual observation vector with itself. It gives an indication for the model fit (see Figure 3).



Evaluation

The squared prediction error (SPE) is a measure for the model fit. Observations above the 95% limit are considered outliers, those above the 99% limit are considered strong outliers.

Standardized Observed Samples and Standardized Predicted Samples

o = observed, p = predicted

#	SPE	Record	Number of Calls	Number of Items	Duration	Price	Number of Participants
0	0.01	12.12.2008, [MCA-At main branchings (Bifurcation)]	o = -0.81; p = -0.72	o = -0.46; p = -0.46	o = -0.62; p = -0.69	o = -0.81; p = -0.82	o = -0.85; p = -0.85
1	0.06	28.05.2008, [PCoA-(Distinct from internal carotid junction)]	o = -0.29; p = -0.35	o = -0.77; p = -0.78	o = -0.35; p = -0.36	o = -0.48; p = -0.28	o = -0.06; p = -0.18
2	0.0	14.04.2008, [MCA-At main branchings (Bifurcation)]	o = -0.55; p = -0.57	o = -0.46; p = -0.46	o = -0.62; p = -0.61	o = -0.48; p = -0.45	o = -0.32; p = -0.33
3	0.03	20.10.2008, [ICA-Posterior Communicating Junction]	o = -1.07; p = -0.95	o = -0.46; p = -0.46	o = -0.9; p = -0.97	o = -0.81; p = -0.89	o = -0.85; p = -0.81
4	0.02	20.10.2008, [BA-Trunk]	o = -0.55; p = -0.45	o = -0.46; p = -0.46	o = -0.35; p = -0.41	o = -0.48; p = -0.53	o = -0.58; p = -0.57
5	0.03	30.05.2008, [Anterior-Communicating]	o = -0.29; p = -0.17	o = -0.46; p = -0.46	o = -0.07; p = -0.13	o = -0.15; p = -0.25	o = -0.32; p = -0.28
6	0.05	10.03.2008, [BA-Termination (caput)]	o = 1.03; p = 1.07	o = 0.16; p = 0.15	o = 1.31; p = 1.21	o = 0.5; p = 0.66	o = 0.47; p = 0.36
7	0.48	07.11.2014, [ICA-Other]	o = 0.0; p = -0.54	o = -0.46; p = -0.46	o = -0.9; p = -0.52	o = -0.81; p = -0.61	o = -0.58; p = -0.62
8	0.03	10.04.2008, [Anterior-Communicating]	o = 0.76; p = 0.85	o = 0.77; p = 0.78	o = 1.03; p = 1.0	o = 0.5; p = 0.38	o = -0.06; p = 0.01
9	3.64	12.03.2008, [BA-Termination (caput)]	o = 3.65; p = 2.4	o = 0.16; p = 0.14	o = 1.58; p = 2.21	o = 1.81; p = 2.98	o = 3.92; p = 3.35
10	0.02	09.09.2008, [MCA-At main branchings (Bifurcation)]	o = 0.24; p = 0.34	o = -0.46; p = -0.46	o = 0.48; p = 0.42	o = 0.18; p = 0.11	o = -0.06; p = -0.03
11	0.03	09.09.2008, [MCA-At main branchings (Bifurcation)]	o = -0.94; p = -0.81	o = -0.15; p = -0.15	o = -0.76; p = -0.83	o = -0.64; p = -0.74	o = -0.72; p = -0.67

Fig. 3. The squared prediction error information board of the PCA result. The line chart shows that some SPE values are exceeding the 95% limit (orange horizontal line) or even the 99% limit (red horizontal line). The table below illustrates the numeric SPE values for each observation, numbered from 1 to n. According to the limit the record exceeds, the record is colored either orange or red. Further the observed and predicted values for each variable are listed.

Contribution plots Contribution plots are a simple graphical way of investigating the contribution of a variable to the PCA model (see Miller et al. [18]). The aim here is to find the variable(s) that contribute most to unusual values detected in the model and to investigate how these values have been achieved. A variable's influence can either be positive or negative and vary in the strength of its contribution. Commonly, bar charts are used for illustration. In our system, those plots have been realized for score values and the squared prediction error of an observation.

Score contribution plot The score of an observation is a vector whose dimension equals the number of principal components. The individual entries of a score vector are computed as a linear combination of the original variables, i.e., as a weighted sum of the original variables. The terms in this weighted sum are the contributions of the original variables to this observation with respect to the principal component and they can be visualized in a score contribution plot, which looks similar to the SPE contribution plot in Figure 4.

SPE contribution plot Every observation has an SPE value. This SPE value is the sum of squares of the entries in the error vector of the predicted vs. the actual observation vector. The individual summands of the SPE value each depend on an original variable and are called the SPE contributions of the individual original variables. They can be visualized in an SPE contribution plot (see Figure 4).

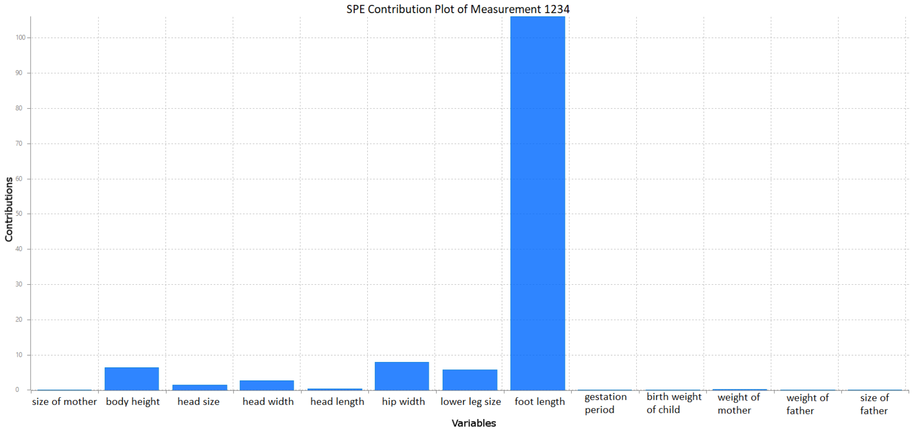


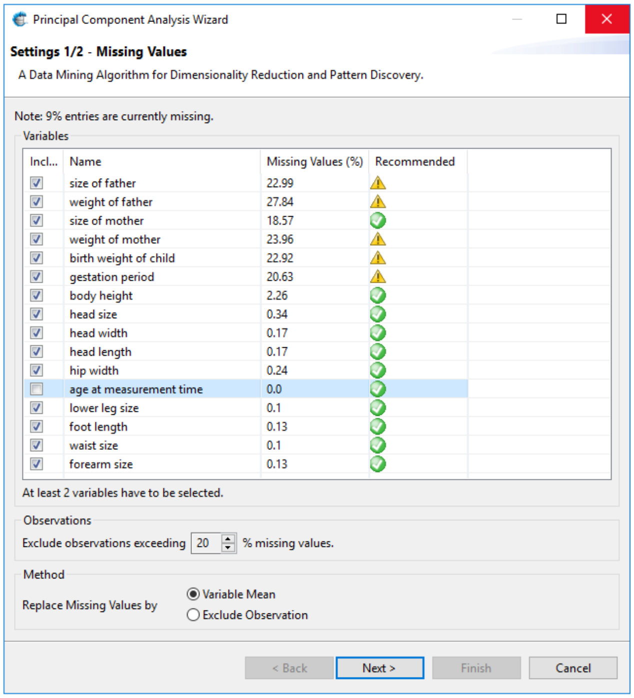
Fig. 4. A squared prediction error contribution plot example of the red marked measurement in Figure 3 on the first principal component. As shown by the figure, the *foot length* contributes the most to this record. When going back to the raw data, this record shows a rather high value in the variable *foot length* compared to the other observations.

It should be noted that the quality criteria listed above include a selection of the most important prerequisites for PCA quality assessment, but may not comprise all relevant prerequisites. If necessary, additional quality criteria may be embedded steadily into the proposed quality measure.

3 Results and Discussion

3.1 Data Set

The MICA (Measurements for Infants, Children, and Adolescents) project started in 2010 with the aim of acquiring detailed demographic and biometric data of children and thus to determine the body surface of children more accurately [19]. In cooperation with the *Kepler University Hospital Linz* about 3200 children aged 0 to 18 were measured by nurses in the outpatient department of the hospital. Those measurements contain more than 30 variables describing biometric variables, such as weight, length and circumference of the child's head, lower legs, hands and feet. Most of the variables are numeric, which have been included in the principal component analysis.



Principal Component Analysis Wizard

Settings 1/2 - Missing Values

A Data Mining Algorithm for Dimensionality Reduction and Pattern Discovery.

Note: 9% entries are currently missing.

Variables

Incl...	Name	Missing Values (%)	Recommended
☑	size of father	22.99	⚠
☑	weight of father	27.84	⚠
☑	size of mother	18.57	✓
☑	weight of mother	23.96	⚠
☑	birth weight of child	22.92	⚠
☑	gestation period	20.63	⚠
☑	body height	2.26	✓
☑	head size	0.34	✓
☑	head width	0.17	✓
☑	head length	0.17	✓
☑	hip width	0.24	✓
□	age at measurement time	0.0	✓
☑	lower leg size	0.1	✓
☑	foot length	0.13	✓
☑	waist size	0.1	✓
☑	forearm size	0.13	✓

At least 2 variables have to be selected.

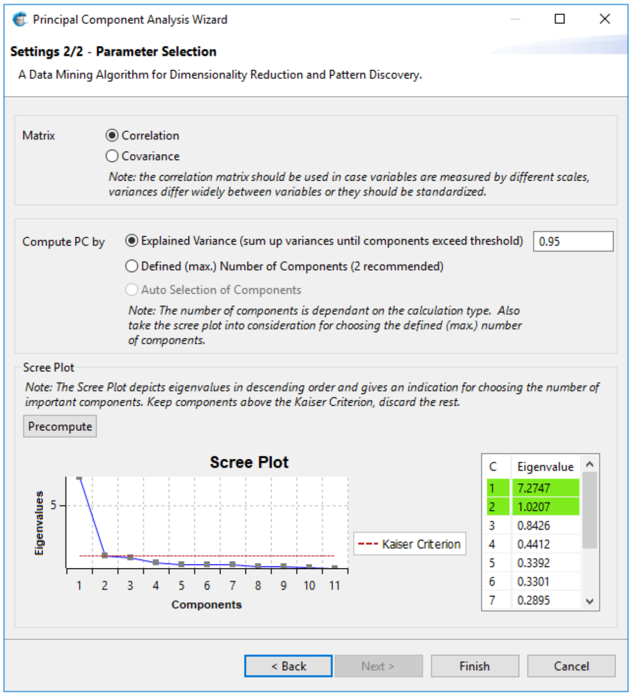
Observations

Exclude observations exceeding 20 % missing values.

Method

Replace Missing Values by ● Variable Mean ○ Exclude Observation

Fig. 5. On the first page of the wizard available numeric variables and the corresponding ratio of null values are shown. Though some of the variables are rated with a warning sign, we consider them to be sufficient for analysis. In this PCA run, missing values will be replaced by the variable mean.



Principal Component Analysis Wizard

Settings 2/2 - Parameter Selection

A Data Mining Algorithm for Dimensionality Reduction and Pattern Discovery.

Matrix ● Correlation ○ Covariance

Note: the correlation matrix should be used in case variables are measured by different scales, variances differ widely between variables or they should be standardized.

Compute PC by ● Explained Variance (sum up variances until components exceed threshold) 0.95

○ Defined (max.) Number of Components (2 recommended)

○ Auto Selection of Components

Note: The number of components is dependant on the calculation type. Also take the scree plot into consideration for choosing the defined (max.) number of components.

Scree Plot

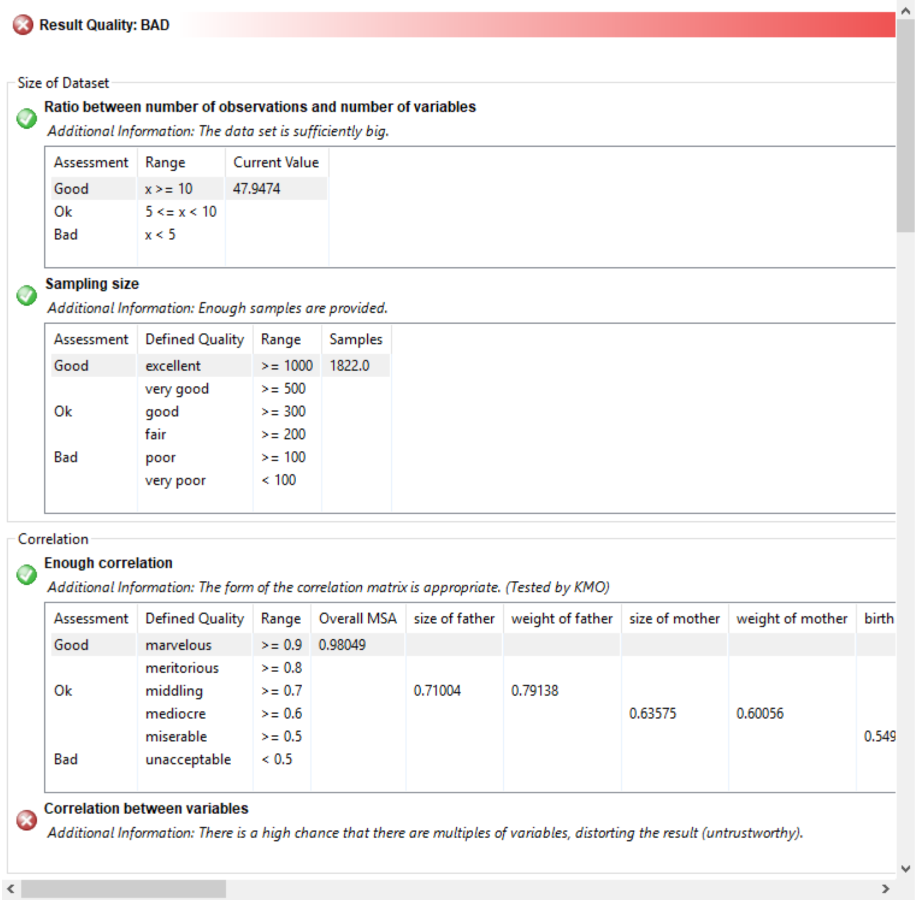
Note: The Scree Plot depicts eigenvalues in descending order and gives an indication for choosing the number of important components. Keep components above the Kaiser Criterion, discard the rest.

C	Eigenvalue
1	7.2747
2	1.0207
3	0.8426
4	0.4412
5	0.3392
6	0.3301
7	0.2895

Fig. 6. The second page of the wizard. The top of the graph illustrates the matrix type selection. Beneath, options for specifying a stopping rule are provided. For support, a scree plot can be precomputed. Within this graph, the Kaiser criterion is illustrated as a second indicator for showing the recommended number of components to be kept. This threshold is visualized by the horizontal line in the chart. In this example, two out of 15 principal components fulfill the requirements, as their eigenvalues are exceeding the threshold (highlighted table entries).

3.2 Approaching Reliable Data Quality

One of the most essential steps in profound data analysis is the preliminary work of data cleansing. Though principal component analysis aims at detecting outliers in data, it is important to remove obviously incorrect data already in advance. Referring to this medical data set obviously erroneous records including values such as negative age at the time of measurement or zero-values in size or weight of the parents have been discarded. Another hot topic is how to handle missing values. Rather than forcible discarding an observation holding at least one single missing value, a softer and more dynamic approach of managing the affected observations is proposed to the user. Scheffer [20] suggests applying single or multiple imputation methods, i.e., to fill in missing values by including means, medians or modes, computed by all known values of the corresponding variable. The default setting in our implementation is the replacement of missing values by the variable mean. If missing values are not to be imputed, the system proposes an alternative configuration – variable and observation exclusion. Both exclusion thresholds can be adapted dynamically by the researcher. However, according to Nelson et al. [21], a higher threshold than 20% missing values leads to losses in performance when applying missing data algorithms (see also Figures 5 and 6).



Result Quality: BAD

Size of Dataset

Ratio between number of observations and number of variables

Additional Information: The data set is sufficiently big.

Assessment	Range	Current Value
Good	x >= 10	47.9474
Ok	5 <= x < 10	
Bad	x < 5	

Sampling size

Additional Information: Enough samples are provided.

Assessment	Defined Quality	Range	Samples
Good	excellent	>= 1000	1822.0
	very good	>= 500	
Ok	good	>= 300	
	fair	>= 200	
Bad	poor	>= 100	
	very poor	< 100	

Correlation

Enough correlation

Additional Information: The form of the correlation matrix is appropriate. (Tested by KMO)

Assessment	Defined Quality	Range	Overall MSA	size of father	weight of father	size of mother	weight of mother	birth
Good	marvelous	>= 0.9	0.98049					
	meritorious	>= 0.8						
Ok	middling	>= 0.7		0.71004	0.79138			
	mediocre	>= 0.6				0.63575	0.60056	
	miserable	>= 0.5						0.549
Bad	unacceptable	< 0.5						

Correlation between variables

Additional Information: There is a high chance that there are multiples of variables, distorting the result (untrustworthy).

Fig. 7. Excerpt of the detailed quality description listing shown to the user. The overall quality is computed as *bad* (red highlighting on the top of the figure). Assessment scheme classification (*good, ok, bad)* of the actual computed key numbers are highlighted in light gray. As an additional information a short and comprehensible description is provided to the user.

Regarding the medical data set, a vast majority of the variables describe measurements of the children. As a matter of fact, some parts of the body grow more or less at the same rate – resulting in similar proportions and therefore multiples in their correlation.

The aim of this evaluation primarily is to showcase the practical use of the quality measure, to illustrate how to derive new hypotheses from the data, and to compare and validate them with existing knowledge, rather than acquiring new medical knowledge or testing hypothesis on a confidence level.

In a first PCA run, including all of the measurement variables, the correlation quality criterion of the quality measure indicated untrustworthy results (see Figure 7). After revising the input variables for achieving reasonably good correlation values, i.e., the occurrence of multiples of variables is very unlikely, the following 15 variables have been included in analysis (see guided parameter selection in Figure 5):

- parent related information such as *the weight and height of the child's mother and father,*
- *gestation period*, and
- child/adolescent related information such as *birth weight, head width, head length, head size, hip width, waist size, forearm size, body height, lower leg size* as well as the *foot length.*

In order to facilitate the application of advanced data mining methods for domain experts, data preparation and parameter selection is intended to be automated as far as possible. In case of PCA, this means choosing the right data matrix as well as an adequate number of principal components. Hence, this is necessary for achieving a good fit of the model while keeping as few components as required for simplifying the model. The substantial challenge here lies in finding the optimal number of principal components to keep. However, a great number of commonly used criteria such as significance tests or graphical procedures are available and may be combined [6]. Thus, a selection of those criteria has been integrated in the research infrastructure.

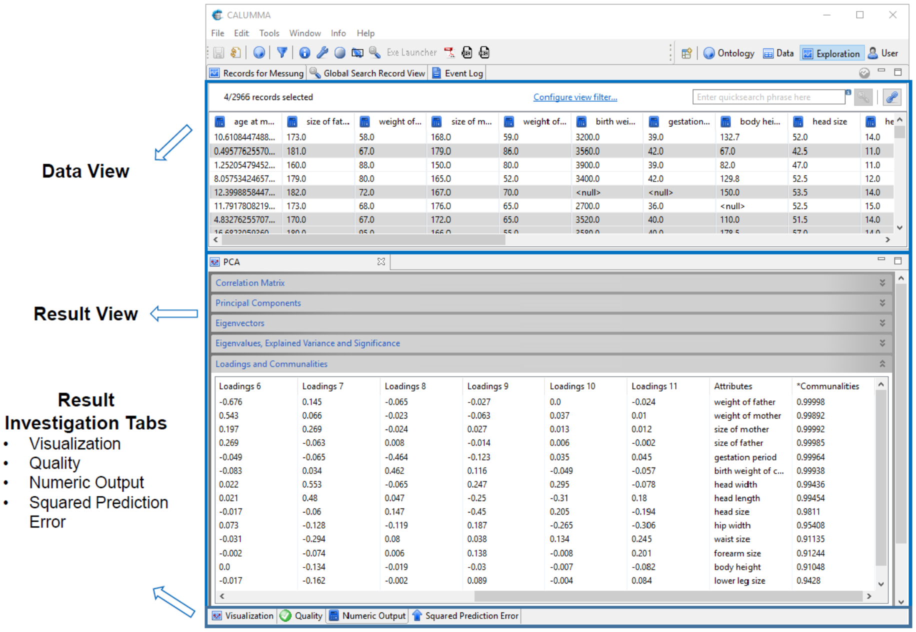
A further graphical approach is provided by the scree test. On the x-axis, all principal components are depicted, whereas the y-axis shows the eigenvalues in descending order. The rationale for this method is that few eigenvalues show particularly high values, accounting for the most variance with a subsequent sudden drop in the eigenvalues. Either the components up to this drop or those with eigenvalues at least as high as a specified threshold (Kaiser criterion [22,13]) are recommended to retain (see Figure 6). This rule should just serve as an approximate value, as it tends to overfactor [23].

In their review, Hayton et al. [23] summarize the benefits from the more accurate factor retention method *parallel analysis.* The basic idea is to compare the eigenvalues of the sample data to those of a number of random generated data (exhibiting the same size and number of variables). What is expected here is, that observed eigenvalues of valid principal components are larger in comparison to the average of the eigenvalues of the parallel components. This method will be provided to users as a supplementary factor recommendation method.

Finally, all necessary input parameters are set to start PCA.

3.3 The Infinite Thirst for Knowledge

Interpreting a PCA output on a numeric base has proven to be complex and difficult for domain experts who are not trained in the fields of mathematics or statistics. When the result is computed, a result package including all relevant information and visualizations of the aforesaid sections is shown to the user.



4/2966 records selected

age at m...	size of fat...	weight of...	size of m...	weight of...	birth wei...	gestation...	body hei...	head size	he
10.6108447488...	173.0	58.0	168.0	59.0	3200.0	39.0	132.7	52.0	14.0
0.49577625570...	181.0	67.0	179.0	86.0	3560.0	42.0	67.0	42.5	11.0
1.25205479452...	160.0	88.0	150.0	80.0	3900.0	39.0	82.0	47.0	11.0
8.05753424657...	179.0	80.0	165.0	52.0	3400.0	42.0	129.8	52.5	12.0
12.3998858447...	182.0	72.0	167.0	70.0	<null>	<null>	150.0	53.5	14.0
11.7917808219...	173.0	68.0	176.0	65.0	2700.0	36.0	<null>	52.5	15.0
4.83276255707...	170.0	67.0	172.0	65.0	3520.0	40.0	110.0	51.5	14.0

Loadings 6	Loadings 7	Loadings 8	Loadings 9	Loadings 10	Loadings 11	Attributes	*Communalities
-0.676	0.145	-0.065	-0.027	0.0	-0.024	weight of father	0.99998
0.543	0.066	-0.023	-0.063	0.037	0.01	weight of mother	0.99892
0.197	0.269	-0.024	0.027	0.013	0.012	size of mother	0.99992
0.269	-0.063	0.008	-0.014	0.006	-0.002	size of father	0.99985
-0.049	-0.065	-0.464	-0.123	0.035	0.045	gestation period	0.99964
-0.083	0.034	0.462	0.116	-0.049	-0.057	birth weight of c...	0.99938
0.022	0.553	-0.065	0.247	0.295	-0.078	head width	0.99436
0.021	0.48	0.047	-0.25	-0.31	0.18	head length	0.99454
-0.017	-0.06	0.147	-0.45	0.205	-0.194	head size	0.9811
0.073	-0.128	-0.119	0.187	-0.265	-0.306	hip width	0.95408
-0.031	-0.294	0.08	0.038	0.134	0.245	waist size	0.91135
-0.002	-0.074	0.006	0.138	-0.008	0.201	forearm size	0.91244
0.0	-0.134	-0.019	-0.03	-0.007	-0.082	body height	0.91048
-0.017	-0.162	-0.002	0.089	-0.004	0.084	lower leg size	0.9428

Fig. 8. Holistic view of the entire system and the integrated result of a PCA run

The holistic view is shown in Figure 8. On top of the graph the data is shown. The bottom section (see Result View) opens subsequently to PCA computation, containing four tabs for result examination. On the present image, the numeric output tab is opened. In the visualization section, scores, loadings, and biplot are shown. The quality tab lists automatically computed quality criteria (see also Figure 7). The last tab shows a view for the squared prediction error (see Figure 3). In order to ensure correct results, all outputs of the implementation have been validated using the R language [24].

In order to facilitate access to interpretations, three types of plots described in Section 1.2 had already been integrated previously in the research infrastructure (see [3]).

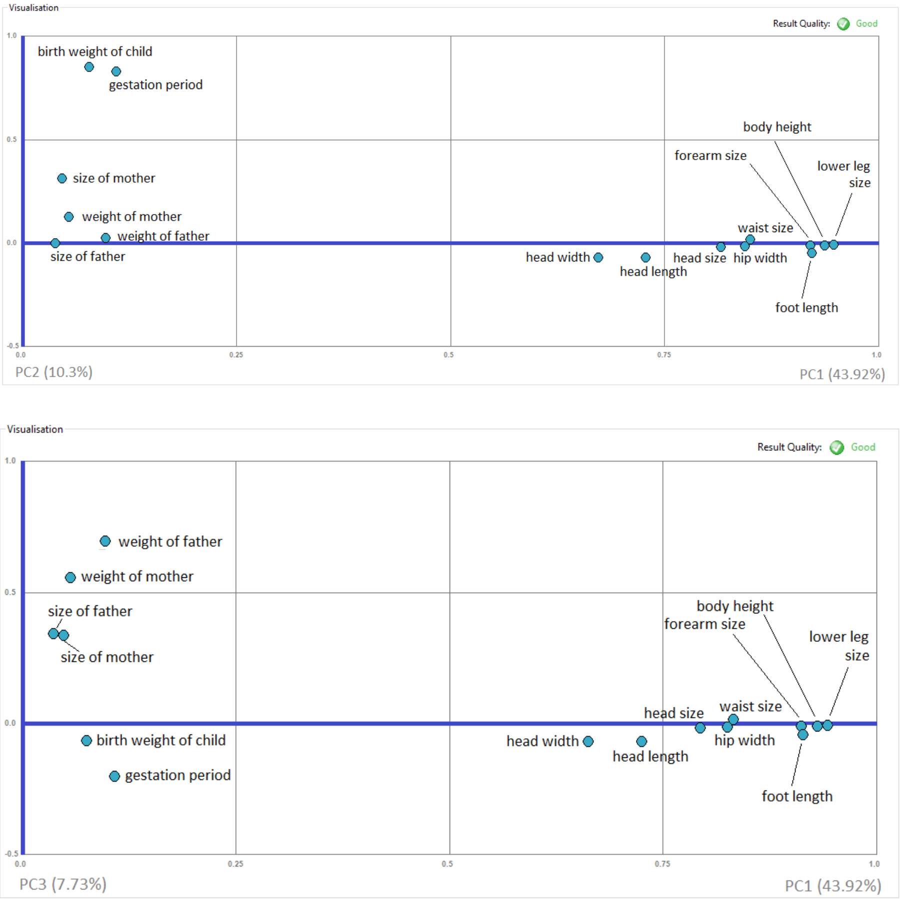


Fig. 9. Loadings plots based on the second PCA result computed in Section 3.2. The first graph illustrates the loadings plot of the first and the second principal component, characterizing the x-axis and y-axis. On the second graph, the x-axis and the y-axis depict the first and the third principal component. The location of the PCA-transformed variables plotted on the graph provides insight into the strength of correlation between the original variables and the influence of a variable on the specific component.

In Figure 9, loadings plots of the current PCA result are depicted. The first graph shows the transformed variables of the data set with the first principal component as the x-axis and the second principal component as the y-axis. Since rather all child related variables are located far from the origin, a strong influence exists by those measurement variables on the first principal component. According to the graph, the child's proportions are correlating strongly with each other, as anticipated. When regarding the second principal component (y-axis), variables contributing to pregnancy and childbirth are described, as they delineate high weights according to this component. Based on our experimental data, due to their proximity, the *birth weight of the child* and the *gestation period* show high positive correlation, i.e., the longer a woman is pregnant, the heavier the child is at birth. It is also apparent that *maternal height* seems to have positive association with the *gestation period*, i.e., that the gestation length is increased for taller women. Literature research actually has shown that this slight influence has already been found in other experiments [25]. Additionally, the third principal component has been evaluated as it gives an indication of the parental influence on the child related data (see Figure 9). According to the positions of *parental body height and weight* divergent to *duration of pregnancy*, influences of maternal and paternal properties are emerging (see also Morrison et al. [26]).

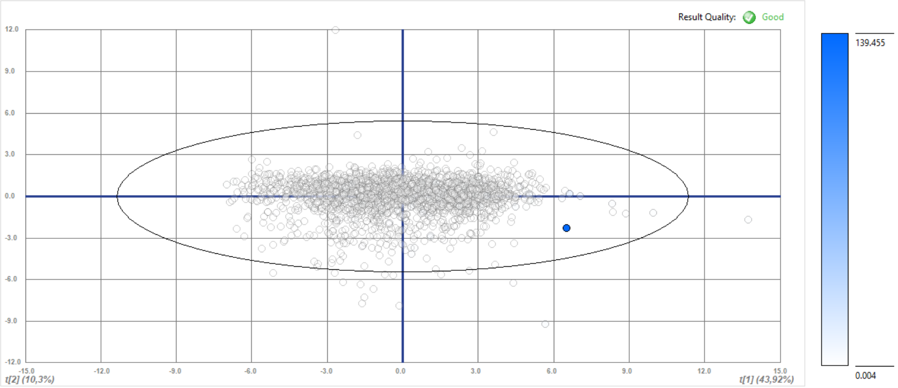


Fig. 10. The score plot (based on the second PCA result computed in Section 3.2) of the first and the second principal component, characterizing the x-axis and y-axis. The color scheme represents the SPE-values of each record. Small deviations (between the actual and the predicted model) are colored white, whereas high differences are highlighted in darker color.

In Figure 10 a score plot of the experimental data is shown. Each point is colored after the squared prediction error. The color represents the distance of the observation to the hyperplane – the darker a point, the higher the distance. In this case, the specified observation is badly explained by the model and is an indication for a potentially corrupted record. It might also be worth to take a closer look at this specific record and to find out why it differs so greatly from the remaining records. Generally, the interactive implementation of the charts provides the connection from drawn points selected in the visualization to the raw data. The corresponding records are then highlighted in the data view (for an example of the data view see Figure 8). Further investigations of the experimental data showed the expected outcome and how to interpret PCA results (see Figure 11). Positive correlations between the growth of included body regions and the stages of age are apparent in each of the plots.

Recall that projections by PCA are only applicable to numeric variables. In case other data types such as categorical variables should be incorporated into analysis, scores can be colored in the corresponding colors and therefore used, e.g., for cluster analysis. Further analysis on comparing the behavior of various subgroups follows the same procedure as introduced, by using subsets of interest based e.g. on gender, ethnicity, or other characteristics. However, it is important to bear in mind the possible bias in these responses as well as the fact that PCA merely can find linear patterns, it is not applicable for detecting non-linear relationships.

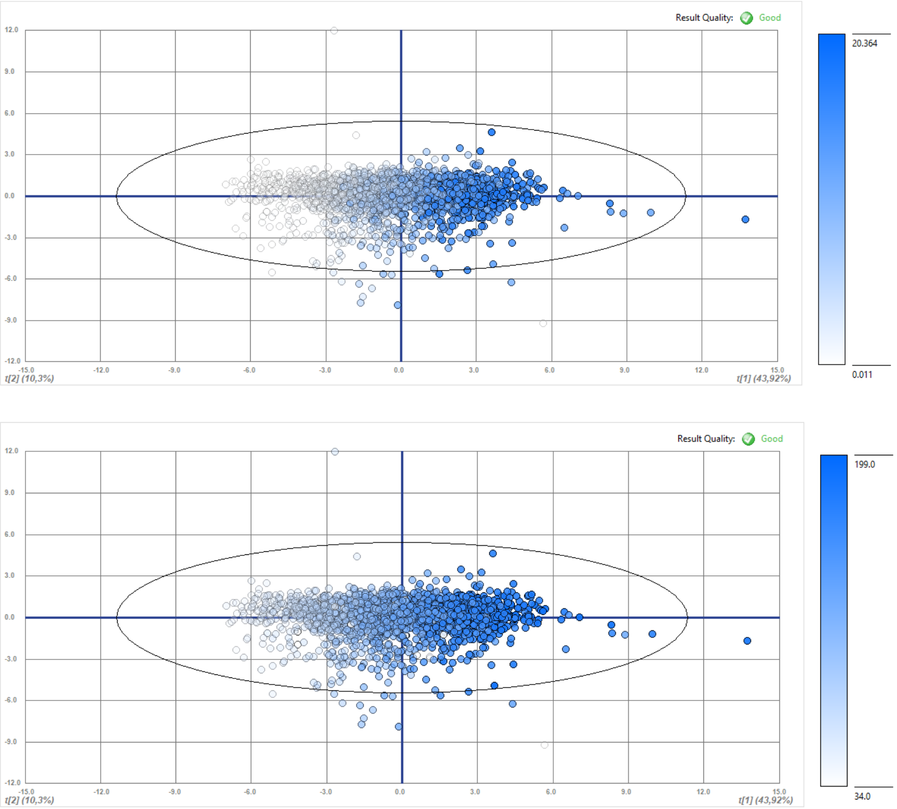


Fig. 11. Age graded score plot (first graph) and measurement graded score plot (second graph). The points are colored from white to blue, whereby the white points on the first graph represent children measured in early years, bluish points represent older test persons. On the second graph white points depict smaller children, the bluish points depict taller children.

4 Concluding Remarks and Future Work

We integrated PCA with an augmented guidance system into an ontology-guided research platform – starting with the control for appropriate input variables and reasonable hyperparameters, going ahead with informing the user about the reliability of the received output and recommendations on how to approach a more trustworthy result, and providing interactive visualizations and result-related information for investigating data patterns and creating new hypotheses. Technically, we introduced a quality measure for PCA which supports the domain expert in checking if the data have the necessary quality and structure for a meaningful application of PCA, and in improving the data in order to increase its quality, e.g., gather more data, remove essentially duplicate variables, etc. We also showed how this guidance works by applying it to the MICA data set. Overall, it is crucial to reach higher interpretability from data mining techniques, so that humans can understand the path from data to results and – by far more important – the meaning and reliability of the result. Further developments of the linkage between ontologies and probabilistic machine learning is a hot future topic and may lead to profound contributions in terms of explainable AI [27]. Future work includes an usability study with medical doctors to evaluate how good the guidance system works in practice. We also plan to integrate other methods augmented with a guidance system like mixed factor analysis, regression and different clustering methods. Another component to include will be support for the user in the choice of the correct statistical test for a problem, depending on (amongst other) what the user wants to examine, the number of variables or the data type of the chosen variables. Therefore the user will be asked the necessary questions in a wizard in advance and a selection of possible tests will be given as a result.

References

1. J. F. Roddick, P. Fule, and W. J. Graco, "Exploratory Medical Knowledge Discovery: Experiences and Issues," *SIGKDD Explorations Newsletter*, vol. 5, pp. 94–99, July 2003.
2. N. R. Anderson, E. S. Lee, J. S. Brockenbrough, M. E. Minie, S. Fuller, J. Brinkley, and P. Tarczy-Hornoch, "Issues in Biomedical Research Data Management and Analysis: Needs and Barriers," *Journal of the American Medical Informatics Association*, vol. 14, no. 4, pp. 478–488, 2007.
3. S. Wartner, D. Girardi, M. Wiesinger-Widi, J. Trenkler, R. Kleiser, and A. Holzinger, "Ontology-guided Principal Component Analysis: Reaching the Limits of the Doctor-in-the-loop," in *International Conference on Information Technology in Bio-and Medical Informatics*, pp. 22–33, Springer, 2016.
4. D. Girardi, J. Dirnberger, and M. Giretzlehner, "An Ontology-based Clinical Data Warehouse for Scientific Research," *Safety in Health*, vol. 1, no. 1, pp. 1–9, 2015.
5. D. Girardi, J. Küng, R. Kleiser, M. Sonnberger, D. Csillag, J. Trenkler, and A. Holzinger, "Interactive Knowledge Discovery with the Doctor-in-the-loop: a Practical Example of Cerebral Aneurysms Research," *Brain Informatics*, vol. 3, no. 3, pp. 133–143, 2016.

6. J. Jackson, *A User's Guide to Principal Components.* John Wiley & Sons, Inc., 1991.
7. A. Rencher, *Methods of Multivariate Analysis.* Wiley Series in Probability and Statistics, Wiley, 2002.
8. W. Kessler, *Multivariate Data Analysis for Pharma-, Bio- and Process Analytics.* WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2007.
9. J. W. Osborne and A. B. Costello, "Sample Size and Subject to Item Ratio in Principal Components Analysis," *Practical Assessment, Research & Evaluation*, vol. 9, no. 11, p. 8, 2004.
10. R. Beaumont, *An Introduction to Principal Component Analysis & Factor Analysis Using SPSS 19 and R (psych Package)*, April 2012.
11. C. D. Dziuban and E. C. Shirkey, "When is a Correlation Matrix Appropriate for Factor Analysis? Some Decision Rules," *Psychological Bulletin*, vol. 81, no. 6, p. 358, 1974.
12. B. G. Tabachnick, L. S. Fidell, and S. J. Osterlind, *Using Multivariate Statistics.* Allyn and Bacon Boston, 2001.
13. H. F. Kaiser, "A Second Generation Little Jiffy," *Psychometrika*, vol. 35, no. 4, pp. 401–415, 1970.
14. H. F. Kaiser and J. Rice, "Little Jiffy, Mark IV.," *Educational and Psychological Measurement*, 1974.
15. D. A. Jackson and Y. Chen, "Robust Principal Component Analysis and Outlier Detection with Ecological Data," *Environmetrics*, vol. 15, no. 2, pp. 129–139, 2004.
16. D. Kim and S.-K. Kim, "Comparing Patterns of Component Loadings: Principal Component Analysis (PCA) versus Independent Component Analysis (ICA) in Analyzing Multivariate Non-normal Data," *Behavior Research Methods*, vol. 44, pp. 1239–1243, December 2012.
17. H. C. Thode, *Testing for Normality.* CRC Press, 2002.
18. P. Miller, R. E. Swanson, and C. E. Heckler, "Contribution Plots: A Missing Link in Multivariate Quality Control," *Applied Mathematics and Computer Science*, vol. 8, no. 4, pp. 775–792, 1998.
19. S. Thumfart, M. Giretzlehner, J. Höller, M. Ehrenmüller, K. Pfurtscheller, H. Haller, *et al.*, "Proportionally Correct 3D Models of Infants, Children and Adolescents for Precise Burn Size Measurement (187)," *Ann Burns Fire Disasters*, vol. 28, pp. 5–6, 2015.
20. J. Scheffer, "Dealing with Missing Data," *Research Letters in the Information and Mathematical Sciences*, vol. 3, pp. 153–160, 2002.
21. P. R. Nelson, P. A. Taylor, and J. F. MacGregor, "Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations," *Chemometrics and Intelligent Laboratory Systems*, vol. 35, no. 1, pp. 45–65, 1996.
22. H. F. Kaiser, "The Varimax Method of Factor Analysis." Unpublished Doctoral Dissertation, University of California, Berkeley, 1956.
23. Hayton, James C and Allen, David G and Scarpello, Vida, "Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis," *Organizational Research Methods*, vol. 7, no. 2, pp. 191–205, 2004.
24. R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.
25. K. Myklestad, L. J. Vatten, E. B. Magnussen, K. Å. Salvesen, and P. R. Romundstad, "Do Parental Heights Influence Pregnancy Length?: A Population-based Prospective Study, HUNT 2," *BMC Pregnancy and Childbirth*, vol. 13, no. 1, p. 33, 2013.

26. J. Morrison, G. Williams, J. Najman, and M. Andersen, "The Influence of Paternal Height and Weight on Birth-weight," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 31, no. 2, pp. 114–116, 1991.
27. A. Holzinger, P. Kieseberg, E. Weippl, and A. M. Tjoa, "Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 1–8, Springer, 2018.