Relating Principal Component Analysis on Merged Data Sets to a Regression Approach

Michael Meyners[1] & El Mostafa Qannari[2]

[1]*Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany*
michael.meyners@udo.edu
[2]*ENITIAA–INRA, Laboratoire de Statistique appliquée à la caractérisation des aliments, 44072 Nantes, Cedex 03, France, qannari@enitiaa-nantes.fr*

Abstract

A method for calculating a consensus of several data matrices on the same samples using a PCA is based on a mathematical background. We propose a model to describe the data which might be obtained e. g. by means of a free choice profiling or a fixed vocabulary in a sensory profiling framework. A regression approach for this model leads to a Principal Component Analysis on Merged Data sets (PCAMD), which provides a simple method to calculate a consensus from the data. Since we use less restrictions on the variables under investigation, the model is claimed to be more general than the model induced by GPA respectively STATIS, which are widely accepted methods to analyse this kind of data. Furthermore, the PCAMD provides also additional opportunities to compare and interpret assessor performances with respect to the variables of the calculated consensus. An example from a sensory profiling study of cider is provided to illustrate these possibilities.

Keywords: principal component analysis, regression analysis, merged data sets, sensory profiling, assessor performance

Introduction

For calculating a consensus from different data matrices given with respect to the same samples, two different methods are widely accepted, namely Generalized Procrustes Analysis (GPA; cf. Gower 1975, ten Berge 1977) and STATIS (which abbreviates the French expression "Structuration des Tableaux A Trois Indices de la Statistique"; cf. Lavit *et al.* 1994, Schlich 1996). In particular they are often used within sensory profiling applications, thus for interpretation purposes we confine ourselves to this framework. However, the generalisation to other applications is obvious. We assume different assessors which took part in a sensory profiling session of several products. The calculated consensus should appropriately represent the differences between the products under consideration. Up to now, the superiority of neither GPA or STATIS has been shown in general, even though the results of a simulation study as well as some theoretical considerations have been given by Meyners *et al.* (2000). The problems arising with theoretical considerations due to the complicated mathematical background of both methods. Furthermore their algorithms lead to time-consuming programming with computer software. A simpler alternative to calculate a consensus from several data sets on the same products has been given by Kunert and Qannari (1999). They propose to define a supermatrix which includes all individual assessor matrices and to calculate a PCA of this supermatrix. The estimated consensus of dimensionality k is then given by its first k principal components.

In this paper we relate this method to a regression model which will be shown to be more general than the model induced by GPA respectively STATIS. A general relationship between PCA and Regression Analysis has been given by Jong and Kotz (1999). Anyway, it will be shown that in the particular case considered here the least squares approach simplifies to a simple PCA on the Merged Data sets (PCAMD). Thus the method of Kunert and Qannari

(1999) will have a stronger justification. Furthermore, it will be shown how different behaviour of the assessors is represented in the contribution of a particular assessor to the different dimensions of the consensus. These values can also be easily calculated within this method and provide several possibilities for interpretation.

Modelling sensory profiling data

We rely on the notation of Kunert and Qannari (1999), i. e. we assume a sensory profiling with m assessors on n products, which gives us the data matrices $X_1,\ldots, X_m$. Meyners *et al.* (2000) propose a reasonable model for this kind of data implicitly given by GPA respectively STATIS. Under their model assumption the individual assessor matrices arise from an underlying true consensus C (which is obviously unknown) by multiplying it with an isotropic scaling factor α_i and an orthonormal rotation matrix R_i. Furthermore a translation matrix T_i and some independent errors with mean zero, arranged in a matrix F_i, say, are added. Thus we get

$$X_i = \alpha_i C R_i + T_i + F_i .$$

It might be wise to generalise this model to allow for different scaling factors for the different attributes of one assessor, even though neither GPA nor STATIS respect for these differences. That is to say that even if there were no random errors at all in this more general model, usually none of the methods would lead to the true consensus C.

In a first step it is necessary to mention that the data of a sensory profiling is usually pre-treated at least in such a way that the matrices are centred column-wise, i. e. the sum over the products is zero for each attribute within each assessor. This holds for GPA and STATIS as well as for the approach of Kunert and Qannari. It is clear from the model that the translation does not affect the matrices anymore as soon as they are centred. Thus without loss of generality and for convenience of notation we can neglect the translation in the following considerations.

The matrices R_i, $i = 1,..., m$, account for possible confusion of the variables given in C as well as they might represent the use of different linear combinations of these variables by assessor i. However, the constraint of them being rotation matrices, i. e. being orthonormal and hence $R_i R_i^T = I$ where I is the identity matrix, implies that the sum of squares of each column of R_i is equal to 1. Note that the k-th column of R_i accounts for how much the variables of C contribute to the k-th attribute of assessor i. Hence from the constraint it is clear that the overall squared contribution remains the same for all attributes of this assessor, which usually might not be fulfilled in practice. We drop the orthonormality assumption by allowing these matrices to be unconstrained and denote them B_i to distinguish them from the rotation matrices. However, on the opposite we constrain the matrix C to have orthonormal columns, which is much more smaller a constraint than the one we just dropped: We assume the product differences to be expressed in some unnamed but orthogonal variables, which leads us to the orthogonality of the columns. This seems sensible, since we are usually not interested in correlated variables to describe the differences, and in particular if we draw a graphical representation we usually take orthogonal axes. Under this assumption the orthonormality is no additional constraint, since we can simply account for scale differences within the attributes of C by choosing the entries of the matrices B_i appropriate, $i = 1,..., m$. Hence it is also unnecessary to consider different scaling factors for the attributes of an assessor, since these can also be accounted for by B_i.

However, it is well known that some assessors tend to use small ranges of scale, whereas others use a wide range of scale. Neglecting these differences would be misleading since within these methods the latter ones would influence the consensus more than the former ones. Thus we take different isotropic scaling factors α_i into account that represent the general behaviour in using the range of scale. Then the model turns out to be

$$X_i = C\,\alpha_i\,B_i + F_i\,. \qquad (1)$$

To avoid different assessor influences due to different behaviour in using the range of scale, the data matrices should be pre-scaled with the inverse of an estimator of α_i. The estimates proposed by Kunert and Qannari (1999) give a reasonable value since they lead to identical overall variabilities for each assessor. Anyway, their denominator is just intended to allow for comparisons with the scaling factors of GPA and might be replaced by any positive scalar without changing the relative distances of interest between the samples within the consensus. Since we are not interested in comparing the results with those of a GPA, we propose to replace their value T by 1. As it will be shown later on, this leads directly to a straightforward interpretation of the results while comparing the assessor performances. Thus we propose to estimate the scaling factors to be

$$\hat{\alpha}_i = \sqrt{\mathrm{trace}(X_i^T X_i)} \tag{2}$$

and to pre-scale the respective matrix with the inverse of this value. After this pre-scaling the sum of squares of all entries of the data matrices is equal to 1 for all assessors.

From now on for convenience of notation we assume the matrices to be pre-scaled according to (2), thus denoting the pre-scaled matrices by Y_i (1) simplifies to

$$Y_i = C\,B_i + E_i, \tag{3}$$

where E_i contains the random errors after pre-scaling, which obviously still have mean zero.

Within (3) the dimensions of the matrices are as follows: C is of order (n, q), where q is the number of dimensions in which there are differences between the products and which is unknown, B_i is a (q, p_i)-matrix, where p_i denotes the number of attributes used by assessor i, and E_i is of order (n, p_i). Furthermore we denote $p = \sum_{i=1}^{m} p_i$, the total number of attributes used.

Model (3) seems similar to the well-known linear model. However, unlike the usual linear regression model we have several unknown parameter matrices, namely C and B_i for $i = 1,\ldots, m$. We are rather interested in C, even though the estimation of the different B_i can provide us with useful information upon the individual assessor appearance as it will be shown later on. It has been pointed out that this model might be more applicable than (1) since there are no constraints about the individual matrices whereas we have just a more reasonable constraint for the matrix C.

Note that no within the model no dimensionality constraints are given at all, i. e. in general the consensus C might provide more attributes than given by each of the matrices B_i. In particular, C might theoretically also include attributes that none of the assessors perceives, that is all entries of (the same) particular rows of B_i would be zero for each $i = 1,\ldots, m$. In practice this case may occur if we have chosen non-representative assessors from a superpopulation, i. e. their are non-negligible differences between the products which have not been detected by the panel. Of course we can only describe those product differences that have been perceived and accounted for by the assessors, thus we hope that we found a panel that accounts for all important differences. Furthermore, from a mathematical point of view it is clear that the observed number k of dimensions within C is bounded by $n - 1$ as well as by p, since we will not need more than $n - 1$ dimensions do describe all differences between n products, whereas on the opposite we cannot estimate more than p dimensions from p attributes. In practical applications we hope to find an estimator for C with a rather small number of attributes that represent the product differences appropriately, even though our model includes also these rather unfamiliar and hopefully seldom cases.

Estimation of C and B_i

It has been pointed out in the previous section that model (3) contains several unknown data matrices. For estimation purposes we use the well-known least squares criterion, i. e. with the obvious notation we want to minimise the value of

$$\sum_{i=1}^{m} \left\| Y_i - \hat{Y}_i \right\|^2 = \sum_{i=1}^{m} \left\| Y_i - \hat{C}\hat{B}_i \right\|^2 = \sum_{i=1}^{m} \left\| \hat{E}_i \right\|^2 , \qquad (4)$$

where

$$\left\| A \right\| = \sqrt{\mathrm{trace}(A^T A)} \qquad (5)$$

is the norm of a matrix A. Due to the fact that it is impossible to estimate those matrices all at once, an iterative algorithm seems to be appropriate. Thus it might be proposed to estimate C in a first step putting in the observed and pre-treated data matrices Y_i for B_i, then to hold the estimated C fixed and estimate the matrices B_i. This should be repeated iterately until convergence occurs. Anyway, we show that no iterations are necessary in practice, since under the assumptions of model (3) the estimate for C is independent of the estimate of B_i but relies only on the matrices Y_i, $i = 1,\ldots, m$.

Let us assume a fixed C that fulfils the constraint of orthonormal columns. Minimisation of (4) is then given as the well-known Gauss-Markov-estimator in a multivariate linear model, i. e.

$$\hat{B}_i = (C^T C)^{-1} C^T Y_i ,$$

which under the constraint $C^T C = I$ simplifies to

$$\hat{B}_i = C^T Y_i \ . \qquad (6)$$

Thus we have

$$\hat{Y}_i = CC^T Y_i$$

and with it

$$\left\|Y_i - \hat{Y}_i\right\|^2 = \left\|Y_i\right\|^2 - 2\,\mathrm{trace}(Y_i^T CC^T Y_i) + \mathrm{trace}(Y_i^T CC^T CC^T Y_i) = \left\|Y_i\right\|^2 - \left\|\hat{Y}_i\right\|^2 .$$

Minimisation of (4) is then equivalent to a maximisation of

$$\sum_{i=1}^{m} \left\|\hat{Y}_i\right\|^2 = \sum_{i=1}^{m} \left\|CC^T Y_i\right\|^2 \tag{7}$$

with respect to *C*. At this stage it should be stated that this is a two-dimensional case of the three-way factor analysis model given by Tucker (1966) and discussed by Brockhoff *et al.* (1996). Anyway, we do not go into details here since this is beyond the scope of this paper. Using (5), (7) can be written as

$$\begin{aligned}
\sum_{i=1}^{m} \left\|CC^T Y_i\right\|^2 &= \sum_{i=1}^{m} \mathrm{trace}(Y_i^T CC^T CC^T Y_i) \\
&= \sum_{i=1}^{m} \mathrm{trace}(Y_i^T CC^T Y_i) \\
&= \sum_{i=1}^{m} \mathrm{trace}(C^T Y_i Y_i^T C) \\
&= \mathrm{trace}\left(C^T \left(\sum_{i=1}^{m} Y_i Y_i^T \right) C \right) \\
&= \mathrm{trace}\left(C^T Y Y^T C\right) ,
\end{aligned} \tag{8}$$

where Y = (Y_1 | Y_2 | … | Y_i) contains the merged original data matrices. It is one of the properties of PCA that the maximisation of this value with respect to k dimensions within C is obtained from the first k eigenvectors of YY^T, i. e. the first k principal components of that matrix. Note that the maximisation does not depend on the values of the matrices B_i. Thus we can estimate C in the first step and after that, if assessor differences are also to be considered, the matrices B_i can be estimated from (6) replacing C by its estimate.

Assessor performance

In practice it seems reasonable to look at the different matrices B_i for the assessors only if we are interested in the contribution of several variables of C to different attributes with respect to particular samples. Usually it is of greater interest to get an expression of the contribution of the assessors to the different principal components, even more if these components have an useful interpretation. We assume $\hat{C} = (\hat{c}_1 \mid \hat{c}_2 \mid \ldots \mid \hat{c}_k)$ to be the k-dimensional approximation of C, where $\hat{c}_j$, j = 1,…, k, are the first k principal components of YY^T. It is also well-known from PCA (cf. Jolliffe 1986) that the j-th eigenvalue λ_k of YY^T is given by

$$\lambda_j = \hat{c}_j^T YY^T \hat{c}_j .$$

Similar to (8), this value can be splitted into m factors which leads to

$$\lambda_j = \sum_{i=1}^{m} \hat{c}_j^T Y_i Y_i^T \hat{c}_j ,$$

and the contribution of assessor i to the k-th principal component is then given by

$$\lambda_j^i = \hat{c}_j^T Y_i Y_i^T \hat{c}_j .$$

These values can be arranged in a table and provide, after different transformations, several interpretations of the assessor performances. In a first step the absolute values should be arranged as proposed in table 1. To simplify the interpretations, we assume that *k* is chosen equal to *m*–1, i. e. we consider all dimensions with nonzero eigenvalues.

assessor	**PC 1**	**PC 2**	…	**PC *k***	Σ
1	λ_1^1	λ_2^1	…	λ_k^1	1
2	λ_1^2	λ_2^2	…	λ_k^2	1
⋮	⋮	⋮		⋮	⋮
m	λ_1^m	λ_2^m	…	λ_k^m	1
Σ	λ_1	λ_2	…	λ_k	*m*

Table 1: Contribution of the assessors to the total variation according to the principal components.

From table 1 it can be seen how much each assessor contributes to the variation with respect to each single dimension. Since the row-wise sum is equal to 1, i. e. all variation of assessor *i* is explained within the principal components given, the entries can be seen as the percentage of variation of assessor *i* explained by the *j*-th principal component. Furthermore the column-wise sum is equal to the respective eigenvalue as it is well-known from PCA. As a matter of fact the eigenvalues add up to *m* when the pre-scaling has been done according to (2). Note that using the pre-scaling proposed by Kunert and Qannari (1999) all values in table 1 are divided by *T*, i. e. to obtain the same table with these interpretations each entry has to be multiplied by *T*.

assessor	**PC 1**	**PC 2**	…	**PC *k***	Σ
1	$\frac{\lambda_1^1}{\lambda_1}$	$\frac{\lambda_2^1}{\lambda_2}$	…	$\frac{\lambda_k^1}{\lambda_k}$	
2	$\frac{\lambda_1^2}{\lambda_1}$	$\frac{\lambda_2^2}{\lambda_2}$	…	$\frac{\lambda_k^2}{\lambda_k}$	
⋮	⋮	⋮		⋮	
M	$\frac{\lambda_1^m}{\lambda_1}$	$\frac{\lambda_2^m}{\lambda_2}$	…	$\frac{\lambda_k^m}{\lambda_k}$	
Σ	1	1	…	1	k

Table 2: Column-wise normalised contribution of the assessors to the total variation according to the principal components.

For table 2 the assessor-contributions have been normalised column-wise such that the columns sum up to 1, i. e. the values account for the percentage the respective principal component is influenced by assessor i. Since these values become quite small as m increases, it might also be useful to multiply all entries with m so that they have mean 1. Furthermore this might simplify the interpretation when we are interested rather in the performance of each assessor than in the percentage of his / her influence on the respective dimension: an assessor with a value larger than 1 for PC j contributes more than the average assessors to this component, whereas an assessor with a value smaller than 1 contributes less. In particular if a principal component is dominated by one assessor, this results in a very large value for the respective assessor. As far as we can see it in this representation the row-wise sums have no useful meaning and are therefore omitted.

The idea to estimate a consensus from sensory profiling data by means of a PCA has been called a "simple alternative" by Kunert and Qannari (1999). Thus it should be mentioned here that our considerations according to the assessor performance respect for this simplicity, since the calculations are easily carried out. It demands only simple programming in any software that provides the possibility to calculate a PCA as well as covariances.

Examples

We consider a sensory profiling of 10 samples of cider that have been assessed by 7 judges according to 10 variables within a fixed vocabulary profiling. We calculated a PCAMD on these data and confine ourselves on the assessor performances, since the interpretation of a product consensus is well-known and in the scope of this paper of minor interest. Table 3 gives the values λ_j^i for i = 1,…, 7 and j = 1,…, 9. As it has been mentioned in the previous section the rows sum up to one whereas the columns sum up to the respective eigenvalues. Due to the rounding used the normalised eigenvalues cumulate not exactly to one.

assessor	**PC 1**	**PC 2**	**PC 3**	**PC 4**	**PC 5**	**PC 6**	**PC 7**	**PC 8**	**PC 9**	**Σ**
1	0.45	0.08	0.13	0.15	0.04	0.04	0.06	0.02	0.02	1
2	0.51	0.14	0.07	0.05	0.05	0.07	0.06	0.03	0.02	1
3	0.38	0.23	0.12	0.05	0.06	0.08	0.03	0.04	0.02	1
4	0.52	0.22	0.07	0.01	0.04	0.03	0.05	0.04	0.02	1
5	0.39	0.15	0.13	0.07	0.05	0.07	0.02	0.06	0.06	1
6	0.46	0.17	0.10	0.08	0.02	0.05	0.03	0.05	0.04	1
7	0.44	0.15	0.08	0.05	0.13	0.02	0.07	0.02	0.04	1
Σ	3.15 (0.45)	1.14 (0.16)	0.70 (0.10)	0.46 (0.07)	0.39 (0.06)	0.36 (0.05)	0.32 (0.05)	0.26 (0.04)	0.22 (0.03)	7 (1.01)

Table 3: Explained variation within the PC’s for the different assessors.

It can be seen that about half of the variation of each assessor is explained by the first principal component, except for assessor 3 an 5 for whom this PC accounts for less than 40%. Assessor 3 has therefore a large value in PC 2, while the remaining variation of assessor 5 is explained by PC 2 and 3. The cumulative variation within the first two components is maximised by assessor 4, i. e. his sample space can be approximated pretty good with the two-dimensional representation from the principal components, whereas assessors 1 and 5 are fairly represented.

assessor	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	Σ
1	1.01	0.50	1.32	2.25	0.80	0.85	1.25	0.46	0.71	
2	1.13	0.84	0.72	0.76	0.92	1.41	1.23	0.82	0.75	
3	0.84	1.42	1.22	0.77	1.03	1.53	0.62	1.02	0.57	
4	1.15	1.36	0.68	0.21	0.78	0.58	1.17	1.01	0.52	
5	0.86	0.92	1.32	1.04	0.88	1.29	0.42	1.72	2.03	
6	1.03	1.06	0.98	1.24	0.34	0.93	0.71	1.37	1.14	
7	0.99	0.90	0.77	0.72	2.27	0.42	1.61	0.59	1.29	
Σ	7	7	7	7	7	7	7	7	7	63

Table 4: Column-wise normalised variation within the PC's for the different assessors.

From table 4 it can bee seen that there are only minor differences in the contribution of the assessors within the first PC, even though assessors 2 and 4 contribute more and 3 and 5 less than the average. Greater differences appear in PC 2 , where assessor 3 (and also 4) contributes quite large, whereas assessor 1 contributes less than the average. Even larger differences can be found in PC 4, 5 and 9, while the smaller eigenvalues and with it the smaller meaning of these components have to be taken into account, i. e. we should avoid any overinterpretation of the values.

To conclude, we might assume assessor 4 to be represented quite well in the first dimensions of the overall consensus and thus to agree well with this consensus, while the assessors 1 and 5 seem to agree less with it, which leads to a small representation within the first two PC's.

Discussion

The simple alternative to GPA proposed by Kunert and Qannari (1999) has been mathematically justified by finding an appropriate model for sensory profiling data. Estimating the unknown parameters within this model leads directly to a very simple method for calculating an overall consensus from several data matrices. Furthermore it gives also a simple opportunity to compare the assessor performances in the panel by calculating the contribution of each assessor to the dimensions of the consensus. The example shows that this might lead to useful conclusions among the assessor behaviour.

Acknowledgement

This paper was written while the first author visited the ENITIAA Nantes. He wishes to thank the people at the ENITIAA for their hospitality and the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity for multivariate data structures") for the financial support of this work.

References

BROCKHOFF, P. M., HIRST, D. and NÆS, T. (1996): Analysing Individual Profiles by Three Way Factor Analysis. *In: Næs, T. and Risvik, E. (ed.): Multivariate Analysis of Data in Sensory Science, 307-342.*

GOWER, J. C. (1975): Generalized Procrustes Analysis. *Psychometrika 40 (1), 33-51.*

JOLLIFFE, I. T. (1986): Principal Component Analysis. New York: Springer-Verlag.

JONG, J.-C. and KOTZ, S. (1999): On a Relation between Principal Component Analysis and Regression Analysis. *The American Statistician 53 (4), 349-351.*

KUNERT, J. and QANNARI, E. M. (1999): A simple Alternative to Generalized Procrustes Analysis: Application to Sensory Profiling Data. *Journal of Sensory Studies 14, 197-208.*

LAVIT, C., ESCOUFIER, Y., SABATIER, R. and TRAISSAC, P. (1994): The ACT (STATIS Method). *Computational Statistics and Data Analysis 18, 97-119.*

MEYNERS, M., KUNERT, J. and QANNARI, E. M. (2000): Comparing Generalized Procrustes Analysis and STATIS. *Food Quality and Preference 11 (1-2), 77-83.*

SCHLICH, P. (1996): Defining and Validating Assessor Compromises about Product Distances and Attribute Correlations. *In: Næs, T. and Risvik, E. (ed.): Multivariate Analysis of Data in Sensory Science, 259-306.*

TEN BERGE, J. M. F. (1977): Orthogonal Procrustes Rotation for Two or More Matrices. *Psychometrika 42 (2), 267-276.*

TUCKER, L. R. (1966): Some Mathematical Notes on Three-Mode Factor Analysis. *Psychometrika, 31, 279-311.*