COMPARING GENERALIZED PROCRUSTES ANALYSIS AND STATIS

M. Meyners,[a] J. Kunert[b] & E. M. Qannari[c]

[a] Fachbereich Statistik, University of Dortmund, D-44221 Dortmund, Germany
E-mail: meyners@amadeus.statistik.uni-dortmund.de
[b] Fachbereich Statistik, University of Dortmund, D-44221 Dortmund, Germany
E-mail: kunert@amadeus.statistik.uni-dortmund.de
[c] ENITIAA/INRA, Unité de Statistique Appliquée à la Caractérisation des Aliments,
Rue de la Géraudière B. P. 82225, F-44322 Nantes, Cedex03, France
E-mail: qannari@enitiaa-nantes.fr

Abstract

We consider a model for sensory profiling data including translation, rotation and scaling. We compare two methods to calculate an overall consensus from several data matrices: GPA and STATIS. These methods are briefly illustrated and explained under our model. A series of simulations to compare their performance has been carried out. We found significant differences in performance depending on the variance of random errors and on the dimensionality of the true underlying consensus. Therefore we investigated on the dimensionality of the calculated group averages. We found both methods to give too many dimensions compared to the true consensus. This finding is supported by some theoretical considerations. Finally we propose a combined approach which takes advantage of both methods and which gave better results in the simulations.

Keywords: Consensus, Dimensionality, GPA, modified GPA, STATIS, Sensory Profiling

1 Introduction

When analysing sensory profiling data several problems unknown in other fields occur. In particular there are three main variations: the assessors use different ranges of scale, they might confuse some of the attributes and they have different zeroes (for details see Arnold and Williams, 1986). Several statistical methods have been developed to handle these problems and to calculate a consensus from data matrices of a sensory profiling experiment. This consensus or group average should reflect the true underlying data structure and indicate which products are similar to each other and which ones differ strongly from each other. Two different methods have been widely accepted for the analysis of such data by now. These are Generalized Procrustes Analysis (GPA, see Gower, 1975) and STATIS (which abbreviates the French expression "Structuration des Tableaux A Trois Indices de la Statistique", see Lavit et al., 1994, or Schlich, 1996). Both methods are useful for fixed vocabulary as well as for free choice profiling data. We illustrate a statistical model to explain sensory profiling data in a formal way.

Both methods give the true consensus if we neglect random errors in our model. We deal with the question whether one of the methods leads to generally better results if random errors are taken into consideration. In this case "better" means that the group average of one method is systematically more similar to the true consensus than the result of the other. To do this a measure of similarity has to be defined. Since theoretical investigations on the distribution of the calculated group averages appear to be difficult, we carried out Monte-Carlo simulations.

2 Model assumptions

We assume there is a number of *n* products or samples to be assessed. For simplicity of notation we confine ourselves to experiments with fixed vocabulary with *m* attributes, although at least our theoretical considerations are also valid for free choice profiling data. Further we have *p* assessors each assessing all *n* products. Hence from each assessor we get an (*n*,*m*)-matrix X_i containing his/her assessments. The rows of X_i correspond to the samples and the columns correspond to the attributes. We assume that each product has some true co-ordinates, measured in some ideal attributes. The (*n*,*m*)-matrix that contains these co-ordinates will be denoted by *C*, the underlying true consensus. Without loss of generality this matrix is supposed to have column-sums zero, that means the centre of the products lies in the origin. While testing a sample the assessors are assumed to not perceive exactly the true *C*, but the consensus with some random errors. We assume that these errors are independent for different products and different variables within each assessor and are also independent between the assessors. Further we assume that the errors of assessor *i* are identically normally distributed with mean zero and unknown variance σ_i^2, the variance depending on the assessor. Let the (*n*,*m*)-matrix E_i contain these errors. Then a preliminary model is given by

$$X_i = C + E_i \ . \qquad (1)$$

However, there are more sources of variation occurring in sensory profiling data. First, there is possibly confusion of the variables. For example assessors might mix up bitterness with astringency and vice versa (see Arnold and Williams, 1986), or they might use a linear combination of several variables for what they denote by e. g. sweet. Such linear combinations can be modelled by multiplication of $C + E_i$ from the right by an (orthogonal) rotation matrix R_i. In order to be a rotation matrix R_i has to fulfil the property

$$R_i R_i^T = I_m. \tag{2}$$

To account for different ranges of scale used by the assessors, we multiply the matrices with an isotropic scaling factor λ_i. Without loss of generality we assume $\lambda_i > 0$ for all i=1,…,p, because if we have a negative λ_i then we can replace R_i by (-1) R_i which is still a rotation matrix. Thus our model (1) modifies to

$$X_i = \lambda_i (C + E_i) R_i . \tag{3}$$

Finally we take shifts of scale into consideration, i. e. we assume that assessors use different zeroes. Such translation is modelled by adding a matrix with identical rows, which can be written as $1_n u_i^T$ where u_i might be any vector of length m and 1_n is the n-vector of ones. Thus our final model is

$$X_i = \lambda_i (C + E_i) R_i + 1_n u_i^T . \tag{4}$$

Note that we get a slightly different model if we modify the order in which these variations occur. However, it can easily be shown that the distribution of X_i does not change if we add the random errors after scaling and rotation. This might be the case if we assume the assessors to perceive the true consensus C exactly right but to make mistakes in giving values to their perceptions. In these cases only the variance σ_i^2 has to be chosen appropriately. Even if both kinds of errors are supposed, the distribution of the X_i can be modelled by (4). Equation (4) has been chosen to describe the model because it simplifies the necessary steps in our simulations.

3 Illustration of the methods

This section gives a short illustration of the properties of both methods in terms of our model assumptions. Both methods try to estimate the true consensus C. The idea of GPA is to find the inverse transformations for the ones made by the assessors. In the first step we multiply with a translation matrix ω that centres the data matrices about the origin. After that, we determine $\hat{\lambda}_i$, an estimate of the inverse λ_i^{-1} of λ_i. Then we determine $\hat{R}_i$, an estimate for the inverse R_i^T of R_i. The estimated consensus is given by the arithmetic mean C_G of the transformed data matrices

$$C_G = \frac{1}{p}\sum_{i=1}^{p}\hat{\lambda}_i \omega X_i \hat{R}_i \,, \tag{5}$$

where ω is the matrix that subtracts the column-wise mean.

STATIS uses the fact that the association matrices

$$W_i = \omega X_i X_i^T \omega \tag{6}$$

contain all information about the n-dimensional product differences independent of the number of attributes. Hence it can be applied to free choice profiling data without adding columns of zeros. The main diagonal contains the squared Euclidean distances of the products from the origin, the other entries are proportional to the cosine of the angle between the corresponding products. So for each non-negative definite (n.n.d.) W_i a matrix $\hat{X}_i$ that solves $W_i = \hat{X}_i \hat{X}_i^T$ can be reproduced. This matrix is uniquely determined except for rotations. Since the sum of n.n.d.

matrices is also a n.n.d. matrix, this holds also if we consider a mean of association matrices. Indeed, STATIS calculates

$$W = \sum_{i=1}^{p} u_i W_i \quad \text{with} \quad \sum_{i=1}^{p} u_i = 1 \quad \text{and } u_i \geq 0 \text{ for every } i = 1,\ldots,p.$$

The weights u_i are calculated via the RV-coefficient, which for matrices X and Y is given by

$$\mathrm{RV}(X, Y) = \frac{tr(XX^T YY^T)}{\sqrt{tr(XX^T XX^T)\; tr(YY^T YY^T)}}. \tag{7}$$

The weights are determined in such a way that assessors who judge similar to the others get bigger weights than those who disagree with most of the panel. Then the consensus C_S is calculated from $C_S C_S^T = W$ which can be solved with the help of the singular value decomposition of W. For details see Schlich (1996).

4 Simulations

We compared the performance of GPA and STATIS under model (4) using Monte-Carlo simulations. For the GPA the improved algorithm of ten Berge (1977) has been used. For STATIS we had to decide whether we use the non pre-scaled or the pre-scaled version. Note that if we do not pre-scale the matrices before calculating the mean, then we disregard that the assessors might use different λ_i. Thus it may happen that a poor assessor gets too much influence. For STATIS, the weights to calculate the mean of the association matrices are developed from the RV-coefficients in such a way, that an assessor who agrees well with the others gets a big weight and an assessor who disagrees with most of the others gets a small weight. Since the RV-coefficient is independent from changes of scale, altogether the former assessor might less influence the mean if he uses a narrow scale than the latter one does by using a wide scale. Therefore we decided to use the pre-scaled version (see e. g. Qannari et al., 1997) as this is clearly more appropriate to our model (4).

To compare the two methods, we derived data matrices from a given consensus *C* according to our model (4). First of all we realise that we do not have to simulate all parts of the model. If we consider the algorithms of both methods we notice that as a first step the data matrices are centred around the origin. Therefore the translation does not influence the results, because after this first step the matrices are identical whether we consider the model given in (4) including the translation or the one given in (3) without it. Hence we can neglect the translation and simulate the simpler model. Another essential step in both methods is the scaling step. For STATIS the data matrices are normalized to uniform lengths. After this step the data matrices are independent of the size of the isotropic scaling factors. This holds also for the scaling step of GPA, for which the scaling factors are determined from minimizing an Euclidean distance. The solution of this minimization problem is independent of the scaling factors in the model. Therefore we can drop the isotropic scaling factors in our simulations, leaving besides C just the rotation matrix and the random errors.

Now assume the X_i being simulated according to (1). In the rotation step of the algorithm of GPA, the rotation matrices $\hat{R}_i$ are determined in such a way that

$$\sum_{i=1}^{p}\sum_{j=1}^{i-1}\operatorname{trace}\left(\hat{\lambda}_i X_i Q_i - \hat{\lambda}_j X_j Q_j\right)^T\left(\hat{\lambda}_i X_i Q_i - \hat{\lambda}_j X_j Q_j\right)$$

is minimal if $Q_i = \hat{R}_i$, $i = 1, ..., p$. Now consider $X_i R_i$ instead of X_i. That corresponds to the use of (3) without the scaling factor. Then

$$\sum_{i=1}^{p}\sum_{j=1}^{i-1}\operatorname{trace}\left(\hat{\lambda}_i X_i R_i Q_i - \hat{\lambda}_j X_j R_j Q_j\right)^T\left(\hat{\lambda}_i X_i R_i Q_i - \hat{\lambda}_j X_j R_j Q_j\right)$$

is minimal if $Q_i = \hat{\hat{R}}_i = R_i^T \hat{R}_i$, $i = 1, ..., p$.

Remembering property (2) of orthogonal matrices, then after this first rotation step the data matrices are independent of the rotation matrices from model (3). Therefore these could also be neglected in simulations for GPA. Indeed, they can also be neglected for STATIS: The consensus of STATIS is derived from the association matrices W_i given in (6). From property (2) it is obvious that the association matrices do not depend on the rotation given in the model.

In all, for the simulations we can confine ourselves to the simple model (1). This is the reason why we constructed the model with this order of adding random errors, rotation, scaling and translation. For another order, we would have had to consider a more complicated model than given in (1).

We simulate data matrices from a consensus C according to (1). To cover some typical situations three different matrices have been chosen:

- the first seven attributes of the second assessor from a free choice profiling study among eight different kinds of yoghurt reported by Dijksterhuis and Punter (1990),
- the scores of the first assessor on nine beef carcasses for seven characteristics reported by Gower (1975),
- a matrix containing random numbers from a rectangular distribution over [0,100] with eleven rows (products) and five columns (attributes).

3, 9 and 15 assessors have been considered, that is 3, 9 and 15 matrices were simulated respectively for each observation. Further we considered one or two outliers. Here outliers are understood in such a way that two rows of a matrix have been exchanged. In practice this means that two samples have been confused for the corresponding assessor. Finally, we varied the variance of the random errors and looked at three different types of assessors. We define an assessor being ordinary if his/her error-variance is the same as the overall variance of the entries in C. A good assessor is defined by having an error-variance which is 25 times smaller than that, and a poor assessor has a variance of the errors which is 25 times larger than the ordinary variance.

For the different situations we carried out 1000 repetitions for 3 and 9 assessors and, due to time consuming calculations, only 500 repetitions whenever we had 15 assessors. We varied the number of good respectively poor assessors and the number of outliers. All simulations were carried out for each of the three consensus matrices. We calculated two different estimated group averages for each simulated data set, one with the help of GPA and the other with the help of STATIS. The performance of the methods was judged by comparing the corresponding group averages to the true underlying consensus C. To make the comparison possible, we needed a measure of similarity. Two measures can be derived directly from the methods. The GPA induces an Euclidean distance between the matrices after a Procrustes rotation. This is given by $\text{trace}\left[(C - C_G R_G)(C - C_G R_G)^T\right]$ or, respectively, $\text{trace}\left[(C - C_S R_S)(C - C_S R_S)^T\right]$. Here, R_G and R_S are the orthogonal matrices that rotate C_G respectively C_S in an optimal way to C (Schönemann, 1966). The smaller this distance is the better the matrices correspond to each other. On the other hand, STATIS induces the use of the RV-coefficient given in (7). It can be shown that the RV-coefficient is equivalent to Pearson's correlation coefficient between the association matrices if these are rearranged as vectors. The bigger this value the better is the agreement between the matrices.

To avoid unfair advantages for one of the methods by using the measure of similarity induced by it, we used both measures parallel. In case a method performs better by means of one measure and worse by means of the other, the corresponding observation has been counted as undecided. So for each simulation, we counted one of the methods as performing better if by means of both measures of similarity its result corresponded better to C than the result of the other method.

5 Results

As the main result of the simulations we can point out that GPA performed better than STATIS in significantly more than 50% of the observations. Some of the results are listed in the following tables. As indicated in the tables the total number of repetitions was 1000 respectively 500 and the number of assessors varied from 3 over 9 to 15. Furthermore we considered different numbers of good respectively poor assessors and also of outliers. The last two columns of each table give the number of simulations for which GPA respectively STATIS performed better. If the sum of the last two values of a row is less than the number of repetitions, the missing observations are those in which the two different measures of similarity gave different results, and where we would not decide which method performed better.

(Table 1 about here)

Looking at Table 1, we notice that for this consensus GPA performed significantly better than STATIS in all those cases where not only poor assessors occurred. In particular, this is independent of the number of outliers. Note that STATIS is assumed to give small weights and with them a small influence on the result to assessors with outliers. If all assessors are poor, then the consensus of STATIS could be fitted better to the true consensus than the consensus of GPA. However, in this situation all data matrices consisted basically of random numbers (see also the next section).

Similar results are to be found in Table 2, where results of the simulations for the consensus constructed from Gower's (1975) data are presented.

(Table 2 about here)

Somewhat different results are derived for the random consensus reported in Table 3. Here we observed STATIS to perform better than GPA in most of the simulations. Especially the occurrence of outliers leads to a better performance of STATIS. Here GPA seems to perform relatively well if there are no good assessors and if we have almost equal numbers of ordinary and poor assessors. It also performed slightly better than STATIS if there are only good assessors and no outliers.

(Table 3 about here)

6 Dimensionality

We investigate on the surprising result that in Tables 1 and 2 STATIS performed better when we considered only poor assessors, while in Table 3 GPA performed better when we considered no good assessors and almost equal numbers of ordinary and poor assessors.

What happens if there are only poor assessors? These have been simulated with random errors that have 25 times the variance of the entries of *C*. Then the variance is so big that the random errors completely spoil the underlying consensus, the simulated data matrices almost contain nothing but random numbers. In practice this would be a situation where all assessors cannot perceive the differences between the products and therefore just give random numbers as assessments. It seems that from purely random numbers STATIS creates a group average that can be fitted better to structured consensus as considered in Tables 1 and 2, while GPA seems to create a group average that can be fitted better to random structures as considered in Table 3. However, both cannot have any useful meaning.

Now consider the cases when there are at least some good assessors. How could it be explained that GPA performed better for the more realistic consensus used in Tables 1 and 2, while it generally performed poorer for the consensus constructed from random numbers?

If we turn our attention to the data matrix of random numbers, we expect this one to have a higher dimensionality than the matrices given from true profiling studies (Wakeling et al., 1992). More precisely, in a PCA of the random structure all components explain a similar amount of variance and therefore the singular values are relatively similar. As a measure of dimensionality we might therefore use the variance of the normalized singular values. Here normalized means that the singular values have been multiplied with a constant, such that they add up to 1. We compared the dimensionality of the calculated group averages for both methods with that of the underlying consensus C by means of Monte-Carlo simulations. The higher the variance of the normalized singular values is, the smaller is the dimensionality of the corresponding matrix. For these simulations the same matrix has been used as C as in Table 1. The results are reported in Table 4. The last three columns give the number of observations in which C_G respectively C_S had a higher dimensionality relative to C and (to compare the methods) the number of observations, in which C_G had a higher dimensionality relative to C_S.

We should note that both methods give results that have too many dimensions in comparison to C, which is reasonable because we add random errors that destroy the underlying structure. Furthermore, STATIS has this problem to a larger extent. This should be the reason why STATIS has a weaker performance if the underlying consensus has a small dimensionality. On the other hand, if the dimensionality of the consensus is already high, then the method cannot give a result with many more dimensions and STATIS performs well.

(Table 4 about here)

7 Theoretical considerations

To support our findings in the simulations, we try some theoretical explanations. First we look at the result C_G of GPA, which is given in (5). If we insert our model assumptions (4) we get

$$C_G = \frac{1}{p}\sum_{i=1}^{p}(\hat{\lambda}_i \lambda_i C R_i \hat{R}_i + \hat{\lambda}_i \lambda_i \omega E_i R_i \hat{R}_i).$$

Note that the rows of E_i are identically distributed. Hence the expected value of $\hat{\lambda}_i \lambda_i E_i R_i \hat{R}_i$ is a matrix with identical rows and the operator ω makes it zero. Therefore the mean of C_G is the expectation of $\frac{1}{p}\sum_{i=1}^{p}\hat{\lambda}_i \lambda_i C R_i \hat{R}_i$ (and if the rotation-estimates $\hat{R}_i$ work well, then it is a multiple of the true consensus). If $C = 0$, we also expect C_G being zero. Furthermore the expected dimensionality of each data matrix does not increase and therefore the expected dimensionality of C_G is similar to that of the true consensus C.

If we insert our model (4) into the equation for the association matrices (6), we obtain

$$W_i = \lambda_i^2 \left(CC^T + \omega E_i C^T + C E_i^T \omega + \omega E_i E_i^T \omega \right).$$

The entries of E_i were assumed to have mean zero and variance σ_i^2 and to be independent from each other. So the expectation of W_i is given by $\lambda_i^2 \left(CC^T + m\sigma_i^2 \omega \right)$. Therefore we expect W to have (n-1) non-negligible dimensions, even if C has only one or two. Hence STATIS in general overestimates the dimensionality if the dimension of C is low. In particular, if $C = 0$, we expect W to be a multiple of ω and C_S away from zero.

It should be mentioned that these considerations also hold if we allow dependencies within the rows and only the rows are independent from each other. This would be the case if the assessments for different products are independent as well as the judgements of different assessors, whereas the assessments of the different attributes for one product by one assessor are allowed to be dependent.

8 A combined approach

As an alternative to the use of usual GPA and STATIS we propose a combined approach, a modified version of GPA. It seems obvious from the results shown that in practical situations GPA performs better than STATIS. Nevertheless the weights used by STATIS appear useful. So we tried to combine the algorithm of the GPA with the weights calculated via the RV-coefficient. A simple approach is to weight the matrices equal to the weights calculated in the algorithm of STATIS. Note that for STATIS we calculate a weighted mean of the association matrices (6), that means the weights are derived for the $X_i X_i^T$. Therefore we should use the square-roots of these weights for GPA because here we calculate a mean from the rotated X_i themselves. Several simulations showed, however, that we get even better if we use the original weights instead of their square-roots. So we calculate the weighted mean

$$C_G = \sum_{i=1}^{p} u_i \hat{\lambda}_i \omega X_i \hat{R}_i$$

where u_i are the weights also used to calculate the mean for STATIS. It is true that the original GPA already downweights poor assessors. The proposed variation does this to a much larger extent.

To compare this method with the usual GPA and STATIS we carried out several simulations for the same situations as described above. Without giving the results in detail we can summarize two aspects:

- the modified GPA performs better than STATIS in significantly more than 50% of the simulations,
- the modified GPA also performs even better than GPA in more than half of the simulations.

Hence the modified approach seems to be a useful alternative to both methods used until now.

9 Conclusions

Under reasonable model assumptions for sensory profiling data we compared GPA and STATIS by means of Monte-Carlo simulations. Under these assumptions GPA performed better than STATIS in significantly more than half of the observations. However, STATIS performed better when we considered basically only random numbers as judgements, or when the underlying consensus had high dimensionality and not all assessors were very good. This finding led us to investigate on the dimensionality of the calculated group average relative to that of C. Simulations as well as theoretical considerations showed that both methods should give a result that has too many dimensions. However, STATIS overestimates the dimensionality of C to a larger extent than GPA. With increasing dimensionality of C this problem seems to become less important. The decision whether the use of GPA or of STATIS should be preferred by means of estimating the dimensionality of the true consensus could be the issue for some future research. However, we propose to use a modified version of the GPA to calculate an overall consensus from several data matrices. This approach combines the algorithm of GPA with the weights of STATIS and yielded significantly better results than both other methods.

Acknowledgement

The paper was written while the second author spent his sabbatical at the ENITIAA de Nantes. He wishes to thank the Volkswagen-Foundation for supporting this visit, and the people at ENITIAA for their hospitality.

References

- ARNOLD, G. M. and WILLIAMS, A. A. (1986): The Use of Generalised Procrustes Techniques in Sensory Analysis. *In: Piggott, J. R. (ed.): Statistical Procedures in Food Research, pp. 233-253.*
- DIJKSTERHUIS, G. B. and PUNTER, P. H. (1990): Interpreting Generalized Procrustes Analysis 'Analysis Of Variance' Tables. *Food Quality and Preference 2, pp. 255-265.*
- GOWER, J. C. (1975): Generalized Procrustes Analysis. *Psychometrika 40 (1), pp. 33-51.*
- LAVIT, C., ESCOUFIER, Y., SABATIER, R. and TRAISSAC, P. (1994): The ACT (STATIS Method). *Computational Statistics and Data Analysis 18, pp. 97-119.*
- QANNARI, E. M., COURCOUX, P., LEJEUNE, M. and MAYSTRE, O. (1997): Comparaison de trois Stratégies de Détermination d'un Compromis en Évaluation Sensorielle. *Rev. Statistique Appliquée XLV (1), pp. 61-74.*
- 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, pp. 259-306.*
- SCHÖNEMANN, P. H. (1966): A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika 31 (1), pp. 1-10.*
- TEN BERGE, J. M. F. (1977): Orthogonal Procrustes Rotation for Two or More Matrices. *Psychometrika 42 (2), pp. 267-276.*
- WAKELING, I. N., RAATS, M. M. and MacFIE, H. J. H. (1992): A New Significance Test for Consensus in Generalized Procrustes Analysis. *Journal of Sensory Studies 7, pp. 91-96.*

Table 1: Simulation results for the consensus derived from the data of Dijksterhuis and Punter (1990).

	number of assessors				better performance for	
repetitions	**total**	**good**	**poor**	**outliers**	**GPA**	**STATIS**
1000	3	0	0	0	928	16
1000	3	3	0	0	778	15
1000	3	0	3	0	123	759
1000	9	0	0	1	1000	0
1000	9	9	0	1	353	0
1000	9	0	9	1	290	549
1000	9	3	3	0	954	0
1000	9	3	3	1	906	0
1000	9	3	3	2	881	0
500	15	0	0	1	500	0
500	15	15	0	1	243	0
500	15	0	15	1	210	211
500	15	0	0	2	500	0
500	15	15	0	2	142	0
500	15	0	15	2	217	212

Table 2: Simulation results for the consensus derived from Gower's (1975) data.

	number of assessors				better performance for	
repetitions	**total**	**good**	**poor**	**outliers**	**GPA**	**STATIS**
1000	3	0	0	1	805	70
1000	3	3	0	1	211	138
1000	3	0	3	1	267	571
1000	3	1	1	0	856	7
1000	3	1	1	1	956	0
1000	9	0	0	0	1000	0
1000	9	9	0	0	1000	0
1000	9	0	9	0	459	354
1000	9	4	0	0	1000	0
1000	9	0	4	0	978	1
1000	9	5	4	0	895	0
1000	9	4	0	1	999	0
1000	9	0	4	1	955	7
1000	9	5	4	1	938	0
1000	9	4	0	2	1000	0
1000	9	0	4	2	943	11
1000	9	5	4	2	929	0
500	15	5	5	0	500	0
500	15	5	5	1	500	0
500	15	5	5	2	500	0
500	15	8	0	0	500	0
500	15	0	8	0	497	0
500	15	7	8	0	499	0
500	15	3	3	2	500	0
500	15	9	3	2	497	0
500	15	3	9	2	500	0

Table 3: Simulation results for the data matrix of random numbers.

	number of assessors				better performance for	
repetitions	**total**	**good**	**poor**	**outliers**	**GPA**	**STATIS**
1000	3	3	0	0	421	403
1000	3	0	3	0	245	549
1000	3	3	0	1	38	701
1000	3	0	3	1	254	544
1000	3	1	1	0	69	546
1000	3	1	1	1	78	450
1000	9	9	0	0	477	370
1000	9	9	0	1	12	922
1000	9	9	0	2	9	962
1000	9	4	0	0	8	931
1000	9	0	4	0	446	355
1000	9	5	4	0	0	993
1000	9	4	0	1	3	931
1000	9	0	4	1	414	355
1000	9	5	4	1	1	956
500	15	8	0	0	5	480
500	15	0	8	0	261	147
500	15	7	8	0	0	496
500	15	8	0	1	3	477
500	15	0	8	1	271	149
500	15	7	8	1	1	488
500	15	8	0	2	2	489
500	15	0	8	2	250	149
500	15	7	8	2	1	476
500	15	3	3	0	65	326
500	15	9	3	0	0	499
500	15	3	9	0	56	331
500	15	3	3	1	56	342
500	15	9	3	1	0	499
500	15	3	9	1	40	289
500	15	3	3	2	76	305
500	15	9	3	2	0	496
500	15	3	9	2	41	303

Table 4: Simulation results for the comparison of the dimensionalities.

repetitions	number of assessors: total	good	poor	outliers	higher dimensionality: for C_G relative to C	for C_S relative to C	for C_G relative to C_S
1000	3	0	0	0	1000	1000	0
1000	3	3	0	0	889	990	0
1000	3	0	3	0	1000	1000	0
1000	3	0	0	1	1000	1000	13
1000	3	3	0	1	891	999	72
1000	3	0	3	1	1000	1000	0
1000	9	0	0	0	1000	1000	0
1000	9	9	0	0	966	1000	0
1000	9	0	9	0	1000	1000	0
1000	9	0	0	1	1000	1000	0
1000	9	9	0	1	936	999	17
1000	9	0	9	1	1000	1000	0
1000	9	0	0	2	1000	1000	0
1000	9	9	0	2	967	1000	16
1000	9	0	9	2	1000	1000	0
1000	9	4	0	0	1000	1000	0
1000	9	0	4	0	1000	1000	0
1000	9	5	4	0	1000	1000	0