Methods to analyse sensory profiling data – a comparison

Michael Meyners

Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany
Tel.: +49 (0)231 755 3181 Fax: +49 (0)231 755 3454
E-mail: michael.meyners@udo.edu

Abstract

The analysis of sensory profiling demands skillful statistical methods to account for different variations that are unknown in other statistical appliances. Besides others, these are the different use of the descriptors by the assessors and the different use of the scales. The two most important approaches to cope with such data are given by Generalized Procrustes Analysis (GPA) and STATIS. Recently, for the latter one several variants have been proposed in order to either simplify the calculation or to improve the results. The aim of this paper is to compare these methods with respect to their performance. For this purpose, a model will be stated to describe the outcomes of a sensory profiling study. On the basis of this model, we give a short insight into the ideas of the methods under consideration, and simulations to compare these methods are realised. From those, systematical differences between the methods occur. Finally, a comparison between the methods with respect to the interpretation of the estimated consensuses is given by means of graphically displaying the outcomes. It will be found that the choice of the method is accidental and can be made according to the simplicity of use for each operator.

Introduction

To analyse sensory profiling data, two different methods are mainly used, namely Generalized Procrustes Analysis (GPA, Gower, 1975, ten Berge, 1977) and STATIS (Lavit, Escoufier, Sabatier & Traissac, 1994, Schlich, 1996). To compare the performance of these methods in practical applications, Meyners, Kunert & Qannari (2000) considered a simulation study which showed the advantage of GPA. This result was also supported by some theoretical considerations. From these it was outlined that the consensus of STATIS is too complex and overestimates the number of relevant dimensions in order to explain the differences between the products of interest. Meyners (2001) generalised these theoretical results to other circumstances. In addition, he proposed a correction of the STATIS consensus which effectuated an improvement of the results. Two other versions of STATIS have also been considered: One uses the arithmetic mean of the association matrices instead of a weighted mean (see also Kunert and Qannari, 1999) and hence simplifies the calculation of the consensus. The other one uses the asymptotic weights in case of an increasing number of products. Obviously, this method is not applicable with real data since the asymptotic results depend on the unknown assessor error variances. It was only considered within a simulation study to examine the outcomes of STATIS if the optimal weights were known. The outline of this paper is to give some insight into the ideas of these methods and to compare them with each other. For this, we give some simulation results as well as a graphical comparison with respect to the interpretation of the results.

The model

A model describing the outcomes of a sensory profiling study is essential to compare the methods. Assuming n products and m descriptors, a first proposal has been given by Meyners *et al.* (2000), being

$$X_i = \lambda_i \,(C + E_i)\, R_i + 1_n u_i^T. \qquad (1)$$

Here, X_i is the matrix that contains the judgements of assessor i, $i = 1, ..., p$, and in which the products are arranged in lines and the descriptors in columns. C is the true underlying consensus reflecting the true differences between the products of interest, and E_i contains the random errors of the corresponding assessor. λ_i is a positive isotropic scaling factor that allows a different range of the used scale and R_i a rotation matrix to model the different use of the descriptors by the assessors. The last term represents a translation, i.e. the use of different parts of the scale, and consists of the vector of ones of length n, 1_n, and an arbitrary vector u_i of length m.

Obviously, for all values $\lambda_i > 0$, this model can be written as

$$X_i = (\lambda_i C + F_i)\, R_i + 1_n u_i^T. \qquad (2)$$

with an appropriate matrix F_i. This form allows an additional interpretation, namely that an assessor does not perceive any differences at all but just gives random numbers as judgements. This can be represented by $\lambda_i = 0$, which would lead to identical values for all products in model (1). Therefore the modified model (2) will be used here.

Methods

Five different methods are considered in this paper, namely GPA, STATIS in its original version, a version of STATIS using the arithmetic mean instead of the original weighted mean, a version of STATIS using asymptotic weights as they will be defined later on, and a corrected version of STATIS.

The idea of GPA is to inverse the transformations given in model (1) respectively (2). For this purpose, an optimal scaling factor, a rotation matrix and a translation are estimated for each assessor such that the matrices match each other as well as possible. After this adjustment, a weighted mean is calculated to estimate the true underlying consensus C. We do not go into the well known details here but refer to the literature: For only two matrices a solution is already given by Schönemann (1966), whereas more details about GPA can be found in Gower (1975) and ten Berge (1977). A nice overview with respect to sensory profiling data is also given by Dijksterhuis (1996).

For the original version of STATIS, the so called association matrices $W_i = X_i X_i^T$ are determined. These contain all information about the product differences given by assessor i. From the association matrices a weighted mean is calculated, for which the weights are determined such that an assessor who gives similar results as most of the other ones will get a larger weight than an assessor who does not. This is meant to upweight good assessors, i.e. those with small random errors as they are represented in matrix E_i respectively F_i. The underlying true consensus is then estimated by means of the principal components of the weighted mean. Details about STATIS can be found e.g. in Lavit *et al.* (1994) and Schlich (1996).

Also, three variants of STATIS are considered, all of which use the association matrices. For two of these methods, only different weights are used. The first one uses the arithmetic mean, i.e. the weight $1/p$ for each assessor. Therefore, all assessors are equally weighted without respecting for their reliability in comparison to the other ones. This approach is similar to the one of Kunert and Qannari (1999).

The second one uses asymptotic weights: Meyners (2001) considers the convergence of the weights for a fixed number of assessors while the number of products tends to infinity. These limits can be determined subject to the covariance matrix of the random errors of each assessor and the true consensus C. Of course these matrices are not known in applications, whereas they might be considered for simulations. Details are beyond the scope of this paper, but it occurs that already for small numbers of products the weights of the original STATIS version do not differ too much from the asymptotic values. Nevertheless it seems to be of interest whether the results would improve if these asymptotic and therefore somehow optimal weights were known.

The last version considered here is a corrected version of the STATIS method. Meyners *et al.* (2000) as well as Meyners (2001) show that STATIS overestimates the complexity of the true consensus. This is due to the use of the association matrices which prevents the random errors to nullify one another even if the mean of these errors is zero. Therefore Meyners (2001) proposes a correction of the STATIS-consensus to adjust for this systematic disadvantage. Without going into details, this correction comprises the estimation of the scaling factors and the errors by means of optimal Procrustes rotation and scaling (cf. Schönemann, 1966), from which a weighted mean might be determined. This mean can then be subtracted from the STATIS-consensus before calculating the principal components. Under additional assumptions, this correction has been proven to give an unbiased and consistent estimator of the true consensus.

Comparison by means of simulations

Before we compare the methods of consideration with respect to the interpretation of the results, we want to know whether or not the methods systematically differ at all. Therefore, we simulate some data from a known consensus according to the model given above, and calculate the consensuses of the different methods. These estimators are compared to the true consensus in order to judge which method derives the most reasonable estimator. The simulation of the data matrices is mainly according to the one presented by Meyners *et al.* (2000), but we added some different covariance-structures of the random errors.

In a first step, only the STATIS versions have been considered. A measure of similarity is given by the RV-coefficient between the estimated and the true consensus. The larger the value of RV, the better the matrices match each other. We simulated 1000 repetitions each and compared each pair of methods. For the underlying true consensus, we used the judgements of the first assessor for nine carcasses with respect to seven descriptors as it is reported by Gower (1975). The following tables give the number of simulations in which each method, given in the column, outperformed the ones given in rows. In a first approach we considered nine assessors with a medium error variance and uncorrelated descriptors. The outcomes of the simulations are given in table 1.

method	original	arithmetic	asymptotic	corrected
original STATIS	0	260	260	1000
arithmetic mean	740	0	0	1000
asymptotic weights	740	0	0	1000
corrected version	0	0	0	0

Table 1: Simulation results to compare the different STATIS variants, nine equally well assessors, uncorrelated descriptors, 1000 repetitions.

It can been shown that in case of equal error variance for all assessors, the asymptotic weights are given by $1/p$, i.e. the version using these weights is identical with the one using the arithmetic mean and therefore none of those methods performs better than the other one (cf. Meyners, 2001). Since the original version outperforms these in about three out of four cases, we conclude that the determination of the weights in STATIS is quite reasonable. Finally, from table 1 we find that the corrected version outperforms all other versions by far and therefore gives a reasonable alternative to the other methods.

For table 2, again 1000 simulations were considered for the same data set. In this case, we simulated three good, three medium and three poor assessors according to different error variances. Furthermore, we assume that in applications the errors between descriptors may be correlated with each other and we therefore considered a correlation of $\rho = 0.2$ for each pair of descriptors.

method	original	arithmetic	asymptotic	corrected
original STATIS	0	1	1000	995
arithmetic mean	999	0	1000	998
asymptotic weights	0	0	0	632
corrected version	5	2	368	0

Table 2: Simulation results to compare the different STATIS variants, nine assessors with different error variance, all pairs of descriptors correlated with $\rho = 0.2$, 1000 repetitions.

From table 2, we find again that the corrected version outperforms the other ones, while this holds to a larger extent for the original version and the one using the arithmetic mean. Also, the original version performs better than the one using the arithmetic mean, which supports our interpretation from above. The asymptotic weights differ from each other which is due to the different error variances of the assessors. In this case, the judgements of the good assessors were weighted with 0.202 and those of the medium assessors with 0.091, while the asymptotic weight of the poor assessors is given by 0.040. This results in better estimates for the underlying consensus than given by the other methods known by now, whereas the corrected version still gives better results, albeit the difference it not as large as in table 1 anymore. Note again that, in application, the asymptotic weights are unknown and therefore cannot be used. They are considered only to judge upon the possible performance in case the variances and covariances of the random errors were known.

Many other circumstances have been considered, e.g. with different correlation structures, different numbers and performances of assessors and different true consensuses. In all cases, similar results have been found: The corrected version performs much better than the original version and the one using the arithmetic mean, i.e. better than those methods that are applicable in practice. Mostly, it performed also better than the version using the asymptotic weights, while sometimes the latter one performed equally well. From these simulations we therefore should recommend the use of the corrected version.

Comparison with GPA

By means of simulations and theoretical considerations, Meyners *et al.* (2000) found that GPA outperforms STATIS. Hence it seems that a reasonable alternative is already given by this latter method. However, as stated by Langron and Collins (1985), for $p > 2$ a drawback of GPA is its iterative algorithm with yet unknown convergence properties. Even though the proposed corrected version needs more computational skills than the original version, it still has no need of an iterative algorithm and is therefore easier to compute, in particular for an increasing number of assessors. Therefore, the corrected version of STATIS might prove as a reasonable alternative to GPA if it is not outperformed by the latter one. To judge upon the performance of the corrected version in comparison to GPA, we compared these methods again by means of simulations. In this case, not only the RV-coefficient was used to judge upon the similarity between the estimated and the true consensus, but also the Euclidean distance after an optimal Procrustes rotation and scaling as proposed by Schönemann (1966). The RV-coefficient is induced by STATIS method, while the Euclidean distance is motivated by GPA, therefore each of these criteria might give an advantage to the method from which it is derived. Hence we conclude that one method performs better than the other one if and only if both criteria give the same result. All other simulations will be judged as undecided. More details about the rationale behind this procedure can be found by Meyners *et al.* (2000).

As we did before, once again we considered different circumstances for the data set of Gower (1975). In each circumstance, 1000 repetitions were simulated and the performance of these methods was compared. The number of assessors varied as well as the number of so called outliers, which are assumed to be assessors for whom two products have been confounded. The assessors were of three possible kind: medium assessors are assumed to have an error variance which is equal to the empirical variance within the underlying consensus. Good assessors will then have an error variance of 1/5th from this empirical variance, while poor assessors are assumed to have five times this variance. Note that these values differ from those used by Meyners *et al.* (2000). In the simulations presented here, the random errors between descriptors are assumed to be independent from each other. Table 3 gives the result for several circumstances.

				better performance of	
p	**good**	**poor**	**outlier**	**GPA**	**corr. STATIS**
9	0	0	0	105	591
9	9	0	0	133	379
9	0	9	0	374	457
9	0	0	1	100	640
9	9	0	1	68	584
9	0	9	1	318	489
9	0	0	2	74	674
9	9	0	2	53	652
9	0	9	2	375	463
9	3	3	0	531	200
9	3	3	1	546	223
9	3	3	2	498	270
9	5	4	0	506	179
9	5	4	1	593	196
9	5	4	2	608	193
15	0	0	0	28	756
15	15	0	0	46	455
15	0	15	0	248	600
15	5	5	0	319	300
15	5	5	1	318	329
15	5	5	2	332	345
15	8	0	0	138	423
15	0	8	0	457	302
15	7	8	0	527	175
15	3	3	0	216	435
15	9	3	0	48	388
15	3	9	0	726	127
15	3	3	1	224	429
15	9	3	1	40	469
15	3	9	1	706	148
15	3	3	2	225	460
15	9	3	2	31	465
15	3	9	2	723	148

Table 3: Simulation results to compare GPA and the corrected STATIS version, different assessor constellations, uncorrelated descriptors, 1000 repetitions.

It can be seen that neither method systematically outperforms the other one. Instead, depending on the circumstances, each of the methods gives better results than the other one sometimes. It seems like GPA performs better whenever we have a rather heterogeneous group of assessors (e.g. 7 good and 8 poor assessors out of 15), while the corrected STATIS version performs better for homogeneous assessor groups, e.g. for 15 medium or 15 good assessors. For additional different circumstances and data sets, the results are similar and will therefore not be given here. In all, we conclude that neither method is superior to the other one, while still the iterative algorithm of GPA might be seen as a drawback in case of a large number of assessors.

Graphical comparison of the results

The results of the simulation study showed systematic differences between the methods considered here. With it, the corrected version proposed by Meyners (2001) outperformed the other versions of STATIS by far except for the one using the asymptotic weights which performed equally well. Furthermore, the corrected STATIS version performed as well as GPA does. Even though it is more complicated than the original STATIS version, it is still easier to perform than GPA since it has no need of an iterative algorithm with unknown convergence properties. Hence it might be considered as a reasonable alternative to the latter one.

In this section, we will consider the differences within the methods with respect to practical aspects. For this purpose, we re-analyse different data sets given in the literature later on (cf. Gower 1975, Dijksterhuis and Gower 1991), but begin with an artificial example. We consider a rectangular true consensus of four products in two dimensions, from which the assessments of a couple of judges has been simulated according to the model given earlier in this paper. From these assessors, the consensuses of the different methods mentioned above are calculated and graphically displayed after being matched to each other as good as possible to allow for a better comparison. The graphical representation of the consensuses can be found in figure 1.

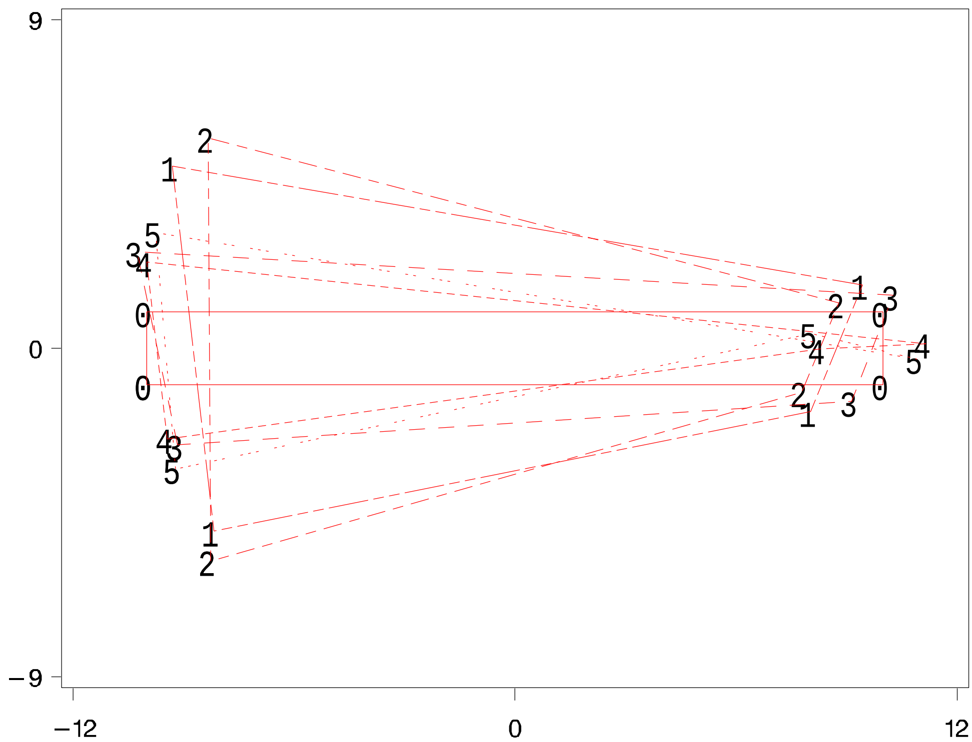

Figure 1: Graphical representation of the consensuses obtained by different methods for an artificial underlying consensus: 0 = theoretical consensus, 1 = original version of STATIS, 2 = STATIS version using the arithmetic mean, 3 = STATIS version using the asymptotic weights, 4 = corrected version of STATIS, 5 = GPA.

In this figure, the edges of the tetragons closest to each other belong to the same artificial product, that is why the "product names" have be omitted. Knowing this, it can be found that the calculated consensuses are quite similar to each other with respect to the interpretation of the results: On each side of figure 1, there are two products that are estimated to be rather similar, while there are large differences to the ones on the other side. Furthermore, the products on the right hand side are estimated to be more similar than the ones on the left hand side. It does not seem not to make much difference which method is used.

Nevertheless, the theoretical results of Meyners *et al.* (2000) and Meyners (2001) can be confirmed by these figures: All methods overestimate the complexity of the true consensus, which is due to the random errors given by the assessors. Also, we find that the consensuses of the original STATIS version and the one using the arithmetic mean are more complex than the true consensus, whereas this is only given to a lower extent for the corrected version, the one using the asymptotic weights and GPA, as can be expected from the references given above.

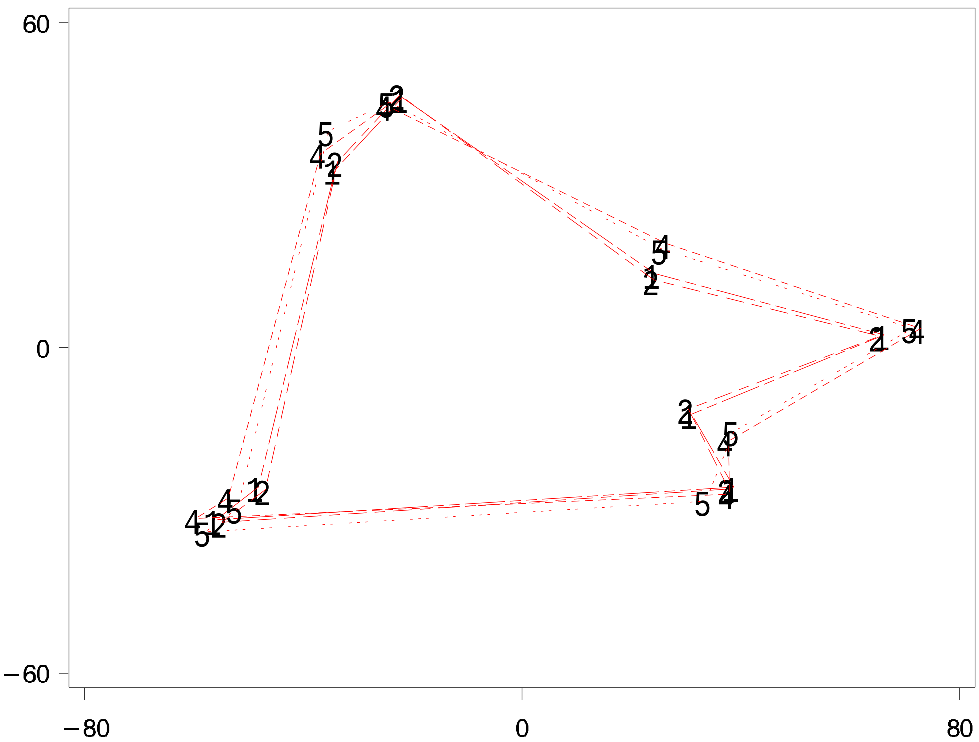

Figure 2: Graphical representation of the consensuses obtained by different methods for the data set presented by Dijksterhuis and Gower (1991): 1 = original version of STATIS, 2 = STATIS version using the arithmetic mean, 4 = corrected version of STATIS, 5 = GPA.

For the data sets from the literature, of course the true consensus is unknown as well as the assessor variances are, and hence also the asymptotic weights cannot be used anymore. Thus we only compare the remaining four consensuses from three STATIS versions and GPA. The respective numbers are therefore omitted in figures 2 and 3. For figure 2, we re-analysed the data set presented by Dijksterhuis and Gower (1991) and determined the different estimators of the true consensus.

Again, the corners of the polygon that are displayed closest together refer to the same product. The most important interpretation is that the estimated consensuses scarcely differ at all. There occur minor differences, which will, however, not lead to different interpretations. In the left and right lower corner, there are two similar products each, as well as in the upper part. In applications, our concern it to determine the products that differ very much from each other and those that do not. For this data set, we would state product differences for exactly the same products irrespective of the method that has been used to calculate the consensus.

Nevertheless, a closer look at figure 2 shows again that the theoretical results mentioned earlier can be supported. Even though the differences are minor, the original version of STATIS as well as the one using the arithmetic mean give less importance to the first and more importance to the second dimension (which are represented by the x- respectively y-axis), i.e. the consensus of these methods is more complex than the one of the corrected version and GPA.

Finally, we consider the data set of Gower (1975). From the representation in figure 3, it seems that there are larger differences between the outcomes of the different methods. Regardless, the interpretation of which products are similar to respectively differ from each other will probably be the same for all methods. It has to be noticed that this data set contains only three assessors, i.e. random differences might cause more disturbance than given in the examples above. In application, sensory profiling studies with only three assessors will scarcely be found.

For the sake of completeness, with a closer look at figure 3 we find again that the estimated consensuses of the original STATIS version as well as the version using the arithmetic mean are of a higher dimensionality than the ones of the corrected version and GPA. The consensuses of the latter ones mainly differ in the first dimension, while the other ones also estimate larger differences in the second dimension as is given e.g. for the products in the mean part of figure 3. This once more confirms the theoretical results given in the literature.

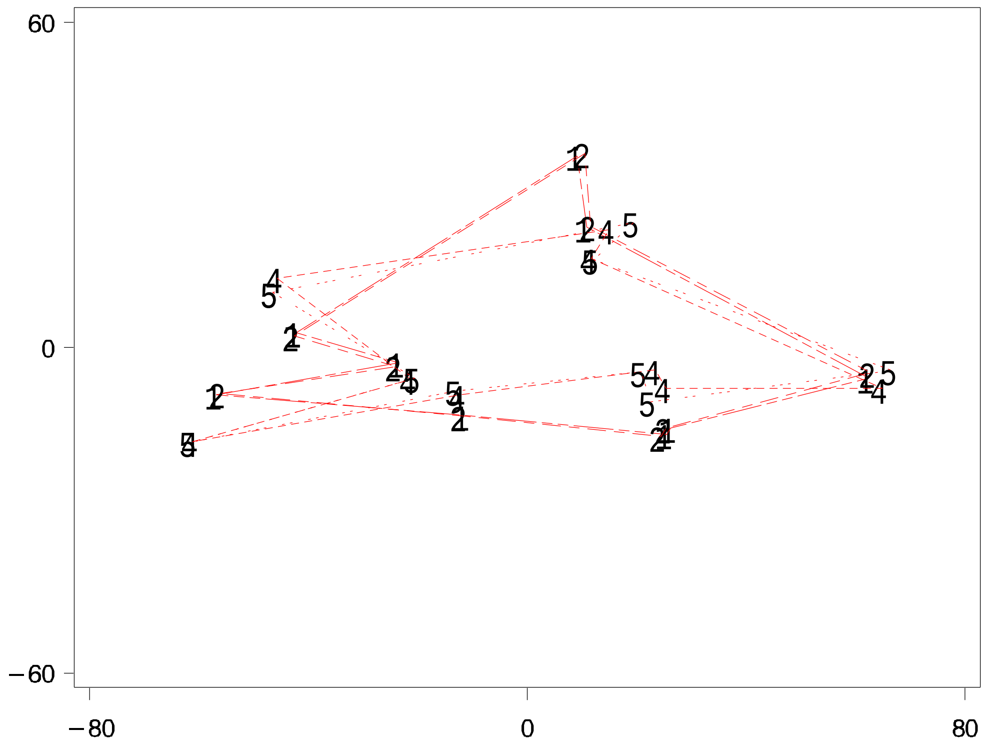

Figure 3: Graphical representation of the consensuses obtained by different methods for the data set presented by Gower (1975): 1 = original version of STATIS, 2 = STATIS version using the arithmetic mean, 4 = corrected version of STATIS, 5 = GPA.

Conclusions

We considered different methods to analyse the outcomes of a sensory profiling study, Generalized Procrustes Analysis and STATIS. From the latter one, also three variants were studied. Two of those are given by another weighting than induced by the original version, namely an arithmetic mean and the asymptotic weights in case of an increasing number of products. The third version uses a correction as it was proposed by Meyners (2001).

The methods under consideration were compared by means of simulations and graphical representations of the outcomes. The results of the simulation study showed systematic differences between the methods considered here. With it, the corrected version proposed by Meyners (2001) outperformed the other versions of STATIS by far, except for the one using the asymptotic weights, which performed equally well in some circumstances. Furthermore, the corrected STATIS version performed as well as GPA does. Even though it is more complicated than the original STATIS version, it may still remain easier to perform than GPA since it has no need of an iterative algorithm with unknown convergence properties. Hence it might be considered as a reasonable alternative to the latter one.

Afterwards, the results of the methods have been compared by means of their graphical representation. We re-analysed some data sets from the literature as well as an artificial data set which allowed to take into account both the underlying true consensus and the asymptotic weights. The resultant figures support that the estimated consensus of the original STATIS version as well as the one using the arithmetic mean is too complex and in particular more complex than the one determined by means of GPA and the corrected STATIS version. This gives another confirmation of the simulation results presented by Meyners *et al.* (2000).

On the other hand side, it can be seen that the graphical representation of the product spread only slightly differs between the methods. The main differences between the products are identically displayed, thus, which is of even more importance, the interpretation of the outcomes will be identical for all methods. Therefore, it might be reasonable to use the simplest version to analyse the data. This might indeed be the version using the arithmetic mean. However, usually different methods are available for the user, from which she or he might choose the one that is the simplest to use for her-/himself. If some software application is available for GPA or STATIS, it seems not to be of much use to investigate in also programming another method. We therefore recommend the use of the method which goes easiest at hand.

Acknowledgement

The author is grateful to the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity for multivartiate data structures") for the financial support of this work.

References

Dijksterhuis, G.B. (1996). *Procrustes Analysis in Sensory Research.* In: T. Næs and E. Risvik (eds.), Multivariate analysis of data in sensory science, 185-219.

Dijksterhuis, G.B. and Gower, J.C. (1991). *The interpretation of generalized procrustes analysis and allied methods.* Food Quality and Preference 3, 67-87.

Lavit, C., Escoufier, Y., Sabatier, R. & Traissac, P. (1994). *The ACT (STATIS method).* Computational Statistics and Data Analysis 8, 97-119.

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

Langron, S.P. and Collins, A.J. (1985). *Perturbation Theory for Generalized Procrustes Analysis.* Journal of the Royal Statistical Society B 47, 277-284.

Meyners, M. (2001). *Statistische Eigenschaften der STATIS-Methode – Propriétés statistiques de la méthode STATIS.* Cuvillier, Göttingen.

Meyners, M., Kunert, J. and Qannari, E.M. (2000). *Comparing Generalized Procrustes Analysis and STATIS.* Food Quality and Preference 11, 77-83.

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.

Schlich, P. (1996). *Defining and validating assessor compromises about product distances and attribute correlations.* In: T. Næs and E. Risvik (eds.), Multivariate analysis of data in sensory science, 259-306.

Schönemann, P.H. (1966). *A Generalized Solution of the Orthogonal Procrustes Problem.* Psychometrika 31, 1-10.

ten Berge, J.M.F. (1977). *Orthogonal Procrustes Rotation for Two or More Matrices.* Psychometrika 42, 267-276.