ELSEVIER

Food Quality and Preference 11 (2000) 85–90

Food Quality and Preference

www.elsevier.com/locate/foodqual

Weighted analysis for missing values in generalized procrustes analysis

Clare Wilkinson*, Maarten Schipper, Tina Leguijt

Agrotechnological Research Institute ATO-DLO, Bornsesteeg 59, PO Box 17, NL-6700 AA Wageningen, The Netherlands

Received 7 September 1998; received in revised form 1 March 1999; accepted 1 March 1999

Abstract

Generalized Procrustes Analysis (GPA), a popular tool in sensory science, is generally carried out on panelist data matrices averaged over replicates. This paper addresses the problem of missing values arising when panelists miss sessions. Because this does not necessarily result in missing values in the final averaged data matrices, a weighted analysis is proposed with weights set proportional to the number of replicates for each panelist product combination. In a simulation study the weighted analysis gives a better match of the rotated panelist matrices (a lower loss) than the unweighted analysis although the resulting average configuration is not significantly closer to the true configuration. The weighted analysis is a straightforward extension to GPA for dealing with missing sessions and offers an improved basis on which to evaluate panelist performance.

1. Introduction

Generalized Procrustes Analysis (GPA) is an increasingly popular tool in sensometrics and is primarily used for dealing with individual differences between panelists. In GPA the data matrices of individual panelists are subjected to rotation and, optionally, transformation and stretching/shrinking to maximize the agreement between the panelists. It uses an iterative algorithm to find rotation and transformation matrices and scaling factors which minimize some measure of the distance between the matrices, the loss function. Details of the algorithm are given in Gower (1975) and, with minor improvements, in ten Berge (1977). The final configurations (the rotated matrices) can then be averaged over panelists to obtain an average, or "consensus" configuration. Consideration of the "distance" of individual panelists from the average configuration leads to information about possible non-conforming panelists. Also the average configuration can be considered an estimate of the unknown "true" configuration.

GPA is a useful method for correcting for certain kinds of differences between panelists, in particular where panelists use different sets of terms to describe the same sensory stimuli. It has the advantage of being a multivariate method, thus dealing with all descriptors and all panelists at once.

The problem of how to deal with missing values in GPA has been considered by several authors in the past. Initially specific patterns of missing values (missing columns or missing rows) were considered. The special case of missing columns is of particular interest to sensory scientists as this situation appears to arise in Free Choice Profiling (FCP). Using this technique, each individual is free to choose which and how many terms he uses, leading to panelist data matrices with unequal numbers of columns. As GPA cannot be directly applied to such matrices, it is useful to consider panelists using fewer than the maximum number of terms as having missing columns. To deal with this situation, Gower (1975) has advocated padding out the matrices with zeroes to achieve equal numbers of columns. With this approach the lower dimensionality of the padded out matrices is preserved (Dijksterhuis & Gower, 1991; Gower, 1995). Alternatively the missing columns are replaced with arbitrary columns which are updated in the course of the algorithm (ten Berge & Knol, 1984). For the general case of randomly dispersed missing values in the data matrices, ten Berge, Kiers, and Commandeur (1993) have recently developed a method. Basically the missing values are imputed, where the imputed value is chosen to minimize a least squares

* Corresponding author.
E-mail address: e.c.wilkinson@ato.dlo.nl (C. Wilkinson).

0950-3293/99/$ - see front matter
PII: S0950-3293(99)00027-0

criterion. This leads to an iterative algorithm where for each panelist in turn the rotation matrix is updated and the missing values imputed until convergence is reached.

The methods described above all deal with situations where there are missing values in the final data set. However in sensory analysis usually the data matrix for each panelist used in GPA is the average of the scores for individual replicates. Thus a missing value for GPA would imply missing values for all replicates. One of the most common causes of missing values in sensory analysis is when a panelist fails to turn up for a session. This leads to a missing replicate for some or all products. However, as long as at least one replicate is present for all products, there will be no missing values for the data matrix used by GPA for that panelist and so the loss of information is hidden.

In this paper a weighted analysis is proposed which takes account of unequal replication. In this way it can deal with the problem of panelists missing sessions. A simulation study is presented which compares the weighted analysis with an unweighted analysis under different experimental circumstances.

2. Method

We consider the case where only rotations are involved (for example, because differences in location and dispersion have been corrected for using a univariate method). Furthermore, we assume missing values can only arise as a result of missing sessions.

Let X_j be the $q \times s$ data matrix for the jth panelist, $j = 1 \ldots p$, containing the scores for the s attributes and q products, averaged over the r replicates. Let W_j be the diagonal $q \times q$ matrix of weights for the jth panelist. The kth element of this matrix, $w_{(j)k}$, $k = 1 \ldots q$, is set equal to r_{jk}, where r_{jk} is the number of replicates produced by panelist j for product k. We wish to find $s \times s$ orthogonal matrices A_j, the rotation matrices, which minimize the loss function, which now incorporates the weights matrices:

$$\text{trace}\left\{\sum_{i<j}\left\{\left(X_iA_i - X_jA_j\right)^T W_iW_j\left(\sum_{i<j} W_iW_j\right)^{-1}\left(X_iA_i - X_jA_j\right)\right\}\right\} \quad (1)$$

Thus distances between points based on many replicates receive more weight than distances between points based on few replicates.

Minimization of the loss function with respect to the rotation matrices is achieved by an iterative algorithm, whereby the rotation matrix for each panelist is updated in turn. For panelist i, the new value for the rotation matrix A_i is obtained from the singular value decomposition of

$$X_i{}^T W_i \sum_{j\neq i} W_jX_jA_j = U\Delta V^T \quad (2)$$

with $A_i = UV^T$.

Further details are given in the Appendix.

The method is illustrated with an example (see Fig. 1). Suppose there are three panelists who each evaluate three products (A, B and C) with respect to two descriptors. Panelists 1 and 2 each have four replicates for the three products. Panelist 3 has four replicates for products A and B but only 1 replicate for product C. The averaged data matrices for the three panelists are

$$\begin{bmatrix} 0 & 0 \\ -5 & -10 \\ +5 & -10 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ +10 & -5 \\ +10 & +5 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ -10 & 0 \\ 0 & 0 \end{bmatrix}$$

and the weights matrices, as defined by the method in this paper, respectively

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The data matrices are shown in Fig. 1. Here a_1 is the point representing the averaged values for panelist 1 given to product A for the two descriptors, b_1 the point for panelist 1 and product B and so on. It can be seen that for panelist 3, in contrast to panelists 1 and 2, the average values for product C (based on one replicate) are identical to those for product A with respect to the two descriptors.

For GPA, rotation matrices need to be found which minimize the sum of the distances between the pairs of points $\{a_1, a_2\}$, $\{a_1, a_3\}$, $\{a_2, a_3\}$, $\{b_1, b_2\}$ and so on. In the weighted analysis the distances $\{c_1, c_3\}$ and $\{c_2, c_3\}$ receive less weight than the distance $\{c_1, c_2\}$, whereas in the unweighted analysis all three distances have equal influence. This leads to different solutions, as can be seen in Fig. 2. Here are shown the average, or "consensus", configurations after GPA for the weighted analysis ($\{a^*_w, b^*_w, c^*_w\}$) and the unweighted analysis ($\{a^*, b^*, c^*\}$). The position of product C in the unweighted analysis has moved closer to product A compared to the weighted analysis. This reflects the greater influence of panelist 3 in the unweighted analysis, whose single replicate score for product C is identical to their average score for product A.

3. Simulations

It can be expected that the added benefit of using a weighted analysis will vary depending on the severity of the impact of the missing values. In order to investigate this further, a small simulation study was carried out.

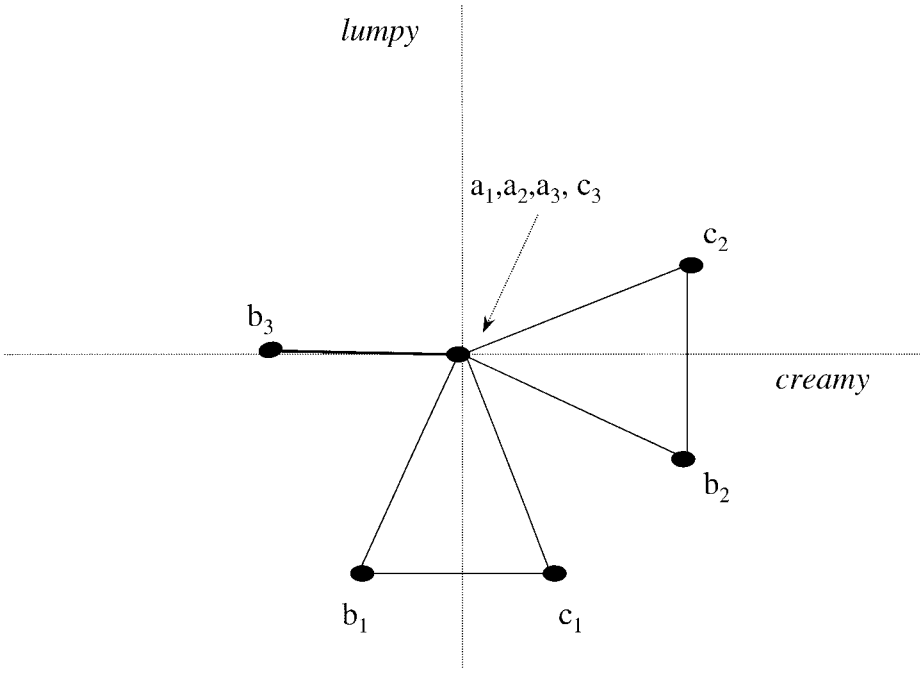

Fig. 1. A representation of the averaged data matrices (artificial data) for three panelists (1, 2 and 3) and three products (a, b and c) for two descriptors (creamy and lumpy).

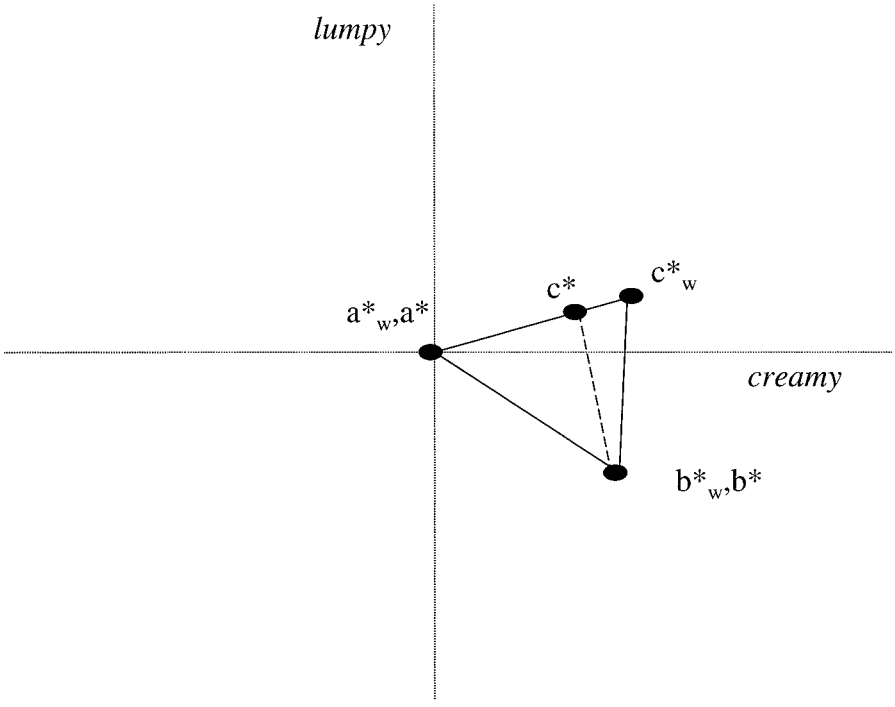

Fig. 2. The agreement or average configuration after GPA on the data matrices shown in Fig. 1, with weights (a^*_w, b^*_w, c^*_w) and without weights (a^*, b^*, c^*).

For this study, 10 "underlying structures" are generated, each consisting of a known "consensus" configuration C0 of six products for three descriptors, and of known rotation matrices for the panelists R_i. The 6×3 matrix C0 is constructed by sampling from the uniform $[-1,+1]$ distribution. The matrices R_i are random orthonormal matrices. From each underlying structure 16 data sets are constructed reflecting different experimental

scenarios. Specifically the following parameters are varied according to a full factorial design:

- the number of products per session; 6 or 3,
- the number of panelists; 3 or 6,
- the number of replicates: 2 or 4
- the probability for each panelist of missing a session: 10 or 40%.

For each data set random error is added, sampled from the normal distribution $N(0,0.05)$. The data sets are then averaged over replicates and GPA is applied, both with and without weights.

To compare the performance, the values for loss and inaccuracy are measured analogous to ten Berge et al. (1993). Loss is simply the final value of the loss function. To measure inaccuracy, the estimated average "consensus" configuration after GPA is rotated to a matched space with the true configuration C0 (this is because a GPA solution is unique except for a common rotation factor) and the distance between the two configurations is calculated.

The simulation data on loss and inaccuracy were analysed using an ANOVA. Some of the results are presented in Tables 1–3. In Table 1 the average value of the loss function and the inaccuracy after weighted analysis is compared with the values after unweighted analysis. In Table 2 the average effects can be seen of the different experimental parameters on the loss function, both overall and when a weighted as opposed to an unweighted analysis is used. Table 3 shows the average value for inaccuracy from a weighted and unweighted analysis and for different values of the experimental parameters (regardless of weighting).

It can be seen that a weighted analysis leads on average to a lower value for the loss function (Table 1). Also, Table 2 shows that the relative benefit of the weighted analysis is greater (if not significantly so) when there are fewer replicates, when the probability of missing a session is higher, and when the number of products evaluated in a session is higher (thus increasing the loss of information when a session is missed). The relative benefit of the weighted analysis is also greater when there are six panelists instead of three. A comparison of the mean values for inaccuracy after a weighted or an unweighted analysis (Table 1) shows a negligible effect of using weights. In other words, in this simulation

Table 1
The overall effect of using a weighted analysis on loss and inaccuracy in the simulation study

	Mean loss function value	Mean inaccuracy
GPA with weights	1.27	0.54
GPA without weights	1.63	0.55
Significance value of effect	$p=0.031$	$p=0.920$

Table 3
The mean inaccuracy (distance of estimated mean configuration from true mean configuration C0) under different experimental conditions

Experimental condition parameter	Parameter value	Mean inaccuracy	Significance of effect
Number of replicates	2	0.80	$p<0.001$
	4	0.28	
Number of products per session	3	0.45	$p=0.104$
	6	0.63	
Probability of missing a session	10%	0.42	$p=0.023$
	40%	0.66	
Number of panelists	3	0.81	$p<0.001$
	6	0.27	

Table 2
The mean final value for the loss function from the simulation study under different experimental conditions and the effect on this value of using a weighted analysis

Experimental condition parameter	Parameter value	Mean loss	Significance of main effect	Mean loss after GPA with weights	Mean loss after GPA without weights	Significance of interaction
Number of replicates	2	2.03	$p<0.001$	1.72	2.35	$p=0.11$
	4	0.86		0.81	0.91	
Number of products per session	3	1.21	$p=0.006$	1.15	1.28	$p=0.17$
	6	1.68		1.38	1.98	
Probability of missing a session	10%	0.94	$p<0.001$	0.89	1.00	$p=0.14$
	40%	1.95		1.65	2.26	
Number of panelists	3	1.62	$p=0.046$	1.54	1.69	$p=0.20$
	6	1.28		1.00	1.57	

study the weighted analysis appears to be better at matching the individual panelists' data matrices, however the estimated mean configuration obtained after a weighted analysis is not on average any closer to the true configuration than after an unweighted analysis.

Tables 2 and 3 also show the influence of the experimental parameters on loss and inaccuracy in the simulation study. Clearly the number of replicates has a big influence on both loss and inaccuracy. The probability of missing a session affects the loss a great deal but has much less effect on the inaccuracy. This is consistent with the greater improvement shown by the weighted analysis for loss than for inaccuracy. Increasing the number of panelists has relatively more effect on inaccuracy than on loss.

4. Conclusions

The addition of weights is a straightforward extension to GPA which gives a means of taking into account missing replicates arising from panelists missing sessions. In fact the weights can be seen as an allowance for differences in variance of the panelist mean data points. The kth row in the data matrix for panelist j is the mean of r_{jk} replications and so has variance equal to ${r_{jk}}^{-1}$ times the variance of the original data points. Setting the weight $w_{(j)k}$ equal to the number of replicates r_{jk} therefore allocates weight proportional to the variance of the mean data points, under the assumption that the variance of the original data points is constant for all panelists, products and descriptors.

The results of the simulation study suggest that a weighted analysis offers a clear advantage over an unweighted analysis with regard to loss but not with regard to inaccuracy. This suggests that a weighted analysis is particularly useful when the main aim of the GPA is to evaluate individual panelists, for example during panel training. When product testing is carried out with a fully trained panel, the primary aim of GPA is to obtain an estimate of the true but unknown average configuration of the products in the sensory space. Here the emphasis is on a lower inaccuracy, and the simulation study indicates that the influence of experimental parameters such as the number of replicates and panelists may be much greater than the use of an unweighted analysis.

In this paper it is assumed that missing values only arise from missing sessions. Further work needs to be done to also incorporate missing values arising from other causes. In particular, as GPA is most often used for FCP data, there is a need for a method which can deal both with unequal numbers of replicates and with the unequal numbers of columns of FCP data. The use of weights as described in this paper in combination with the method for unequal column numbers proposed by Gower (1975) or that proposed by ten Berge and Knol (1984) needs to be further investigated.

Acknowledgement

We would like to thank Professor J.H.A. Kroeze for his help and advice in preparing this paper.

Appendix

We wish to find $s \times s$ orthogonal matrices A_j (the rotation matrices) which minimize the loss function

$$\operatorname{trace}\left\{\sum_{i<j}\left\{\left(X_iA_i - X_jA_j\right)^T W_iW_j\left(\sum_{i<j} W_iW_j\right)^{-1}\left(X_iA_i - X_jA_j\right)\right\}\right\} \tag{3}$$

To estimate A_i, assuming all other A_j are fixed, we can write the criterion as

$$\operatorname{trace}\sum_{i\neq j}^{p}\left\{(X_iA_i)^T W_iW_j(X_iA_i) + \left(X_jA_j\right)^T W_iW_j\left(X_jA_j\right) - 2(X_iA_i)^T W_iW_j\left(X_jA_j\right)\right\} \tag{4}$$

We therefore need to find A_i such that this criterion is minimized. Now because ${A_j}^TA_j = I$, we can write

$$\begin{aligned}\operatorname{tr}(X_iA_i)^T W_iW_j(X_iA_i) &= \operatorname{tr}(X_iA_i)(X_iA_i)^T W_iW_j \\ &= \operatorname{tr}X_iA_i{A^T}_i{X^T}_i W_iW_j \\ &= \operatorname{tr}X_i{X_i}^T W_iW_j\end{aligned} \tag{5}$$

Now A_i only appears in the third term. So to minimize the criterion we need to find the A_i which maximizes the third term. Rewriting this as

$$\operatorname{tr}\left\{{A_i}^T{X_i}^T W_i\sum_{j\neq i} W_jX_jA_j\right\} \tag{6}$$

then A_i is obtained from the singular value decomposition of

$${X_i}^T W_i\sum_{j\neq i} W_jX_jA_j = U\Delta V^T \tag{7}$$

with $A_i = UV^T$.

Thus the algorithm is as follows:

1. Initialize A_i for all i (e.g. A_i =identity matrix).
2. For each panelist i, $i = 1 \ldots p$, calculate new A_i.
3. Calculate loss function. If it has reduced more than a set tolerance, repeat step 2, otherwise stop.

The average configuration or ''consensus'' configuration is equal to

$$\left(\sum_{i=1}^{p} W_i\right)^{-1} \sum_{i=1}^{p} W_i X_i A_i \tag{8}$$

References

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

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika, 40*(1), 33–51.

Gower, J. C. (1995). Orthogonal and projection procrustes analysis. In W. J. Krzanowski, *Recent advances in descriptive multivariate analysis* (pp. 113–134). Oxford: Oxford University Press.

ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, *42*(2), 267–276.

ten Berge, J. M. F., Kiers, H. A. L., & Commandeur, J. J. F. (1993). Orthogonal procrustes rotation for matrices with missing values. *British Journal of Mathematical and Statistical Psychology, 46*, 119–134.

ten Berge, J. M. F., & Knol, D. L. (1984). Orthogonal rotations to maximal agreement for two or more matrices with different column orders. *Psychometrika, 49*, 49–55.