Received: 18 September 2006, Revised: 4 March 2007, Accepted: 12 March 2007, Published online in Wiley InterScience: 2007

(www.interscience.wiley.com) DOI: 10.1002/cem.1097

Multi-way models for sensory profiling data

Rasmus Bro[a]*, El Mostafa Qannari[b], Henk A. L. Kiers[c], Tormod Næs[d] and Michael Bom Frøst[a]

One of the problems in analyzing sensory profiling data is to handle the systematic individual differences in the assessments from different panelists. It is unavoidable that different persons have, at least to a certain degree, different perceptions of the samples as well as a different understanding of the attributes or of the scales used for quantifying the assessments. Hence, any model attempting to describe sensory profiling data needs to deal with individual differences; either implicitly or explicitly. In this paper, a unifying family of models is proposed based on (i) the assumption that latent variables are appropriate for sensory data, and (ii) that individual differences occur. Based on how individual differences occur, various mathematical models can be constructed, all aiming at modeling simultaneously the sample-specific variation and the panelist-specific variation. The model family includes Principal Component Analysis (PCA) and PARAllel FACtor analysis (PARAFAC). The paper can be viewed as extending the latent variable approach commonly based on PCA to multi-way models that specifically take certain panelist-variations into account. The proposed model family is focused on analyzing data from quantitative descriptive analysis with fixed vocabulary, but it also provides a foundation upon which comparisons, extensions and further developments can be made. An example is given which shows that even for well-working data, models handling individual differences can shed important light on differences between the quality of the data from individual panelists. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: Individual differences, tensor models, PARAFAC

1. INTRODUCTION

The models described in this paper are specifically and primarily aimed at exploring the product space with the key consideration being how to deal with individual differences between panelists. Thus, methods dedicated solely for assessing panelists' performance or for relating sensory data to, e.g. instrumental responses or preference data are not directly considered. Several papers have discussed the nature of individual differences in sensory data [1–4].

For a single attribute, the main differences between assessors can be partitioned into differences in either (i) level of scale, (ii) scaling or sensitivity, (iii) disagreement or (iv) variability [5]. It will be shown how these differences can be handled in latent variable models. The proposed models can be seen as extensions of the use of principal component analysis and related methods, i.e. based on assuming that there is a *latent structure* among the descriptors. The proposed models retain the property of describing the data through latent variables, while taking specific care of individual differences. The use of saliences as used, e.g. in INDSCAL [6,7] are promoted here as a rational way of obtaining estimated parameters pertaining to the quality of the assessments of individual panelists.

The data types discussed are primarily those arising from Quantitative Descriptive Analysis (QDA) with fixed vocabulary profiling. The data will be of the general form: K panelists judging I products with respect to J attributes. The assessments or scorings for the kth panelist can be arranged in a configuration matrix of size $I \times J$ called $\mathbf{X}_k$. The problem is how to determine a condensed set of characteristics of the samples, $\mathbf{T}$ ($I \times F$), where F is to be determined. This is a consequence of the assumption that the J descriptors are representing F ($<J$) underlying latent variables and hence are not independent from each other. The matrix $\mathbf{T}$ represents these F underlying variables in such a way that each column is the magnitude or score of the corresponding latent variables. Rather than scores of J original descriptors, the end result of a latent variable model is hence a score matrix $\mathbf{T}$ with only F variables/columns.

An auxiliary important task is to be able to explore what these condensed characteristics represent in terms of the attributes as well as to what degree the panelists agree upon these characteristics. The matrix $\mathbf{T}$ describes the variation in the samples preferably with the panelist specific variation filtered off.

* Department of Food Science, University of Copenhagen, 1958 Frederiksberg C, Denmark.
E-mail: rb@kvl.dk

a R. Bro, M. B. Frøst
Department of Food Science, University of Copenhagen, 1958 Frederiksberg C, Denmark

b E. M. Qannari
ENITIAA/INRA Unité de Sensométrie et Chimiométrie, Rue de la Géraudière BP 82 225, 44322 NANTES cedex 03, France

c H. A. L. Kiers
Heymans Institute (DPMG), University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

d T. Næs
MATFORSK, Oslovejen 1, 1430 Ås, Norway

Thus, we seek a matrix **T** based upon the K configuration matrices

$$\mathbf{T} = f(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K) \qquad (1)$$

If random Independent and Identically Distributed (iid) variation was the only difference between the panelists, the measurements from each panelist could be considered as a replicate configuration. In that case, the problem could be expressed in terms of the average configuration

$$\bar{\mathbf{X}} = \frac{\sum_{k=1}^{K} \mathbf{X}_k}{K} \qquad (2)$$

as the problem of finding

$$\mathbf{T} = f(\bar{\mathbf{X}}) \qquad (3)$$

When the panelists can be considered to differ only by random variation, **T** is thus a function of the average configuration matrix. A typical way of finding **T** in such cases is by assuming that the attributes are manifest variables expressing the variation in the underlying latent variables, i.e. an underlying basic set of phenomena arising upon sensation. If each latent variable can be approximated as a linear combination of the manifest attributes, a bilinear model can model the data as

$$\bar{\mathbf{X}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (4)$$

Each latent variable is here given by a loading vector, $\mathbf{p}_f$, corresponding to the fth column in the $J \times F$ matrix **P** and a score vector, $\mathbf{t}_f$, corresponding to the fth column in the $I \times F$ matrix **T**. The loading vector $\mathbf{p}_f$ contains J loadings and the score vector of the fth component is simply the weighted sum of all variables, the weights being the elements of $\mathbf{p}_f$. The overall model of the average configuration is the sum of the systematic part, $\mathbf{TP}^{\mathrm{T}}$, plus the unmodeled residuals, **E**. The number of components, F, is not known in advance but must be determined from the data. This bilinear model is the basis for the discussion in this paper and is called the *ideal model* in the following, because it is based on a number of idealized assumptions and because it is ideal in the sense that no panelist-specific contributions occur. Levels or offsets are not included in the ideal model because these are usually assumed panelist-specific. These will be treated later.

In a situation, where the panelists show other individual differences than random, the problem can be considered to estimate **T** of the idealized average configuration given the manifest configurations, $\mathbf{X}_k$, $k = 1, \ldots, K$. Using the average (consensus) configuration will naturally bias the estimated **T** if the panelists do not have the same expectation. A common way of solving this problem is to use a model, that specifically aims at separating the effects of the individual differences between panelists and the common variation as given by **T** and **P**. Specifically we will discuss how to deal with data where the panelists

1. Have (different) levels or offsets on the attribute scales.
2. Have differential scale or sensitivity towards different *latent* variables (e.g. that certain panelists score high with respect to sweetness-related attributes because they perceive such characteristics more intense, but the same panelists do not perceive texture-related descriptors different from the rest of the panel).
3. Have differences in variability/uncertainty.

Systematic disagreements between assessors are not considered explicitly but will show up indirectly as differences in variability or in the residual analysis.

We will outline how the basic bilinear Principal Component Analysis (PCA) model may be accommodated to incorporate the above-mentioned individual differences between panelists. Each case will be treated separately first and only the structural (algebraic) models are considered here. Afterwards, an application is given to highlight some of the possibilities.

2. MODELING INDIVIDUAL DIFFERENCES

Two scenarios for handling different individual differences are given. The models within each scenario are nested in a hierarchical way as described by References [8,9]. This means that for a model of a given complexity, this model will also entail the features of any model of lower complexity within the scenario, hence be able to model the variation of the lower complexity. Thus, only one structure within each scenario must be chosen, namely the one that offers the best compromise between fit and variance. In sensory analysis, this choice of model-type is usually based on methods such as lowest cross-validation error in combination with visual interpretation. More often, though, it is based on experience that a certain type of model is adequate for highlighting important aspects of the empirical data. Such an experience-based choice of model is sensible, especially for sensory data, where individual data sets to be modeled are mostly too small and diverse to enable more rigorous approaches to testing adequacy of several different types of models. In the following, a hierarchy of models are developed, but with the main emphasis on pointing out the possibility of advancing typical PCA models to handle a model structure that take assessor differences into account.

2.1. Scenario A: the panelists have (different) offsets on the attribute scales

A common characteristic of sensory data is that the panelists have different offsets on the attribute scales. This violates the assumption of interval-scale data underlying the ideal model and thus has to be incorporated into the model. Assume that every individual configuration can be modeled as the idealized configuration plus a common offset

$$\mathbf{X}_k = \mathbf{TP}^{\mathrm{T}} + \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{E}_k, \quad k = 1, \ldots, K \qquad (5)$$

where **1** is an I-vector of ones and **m** is J-vector holding in its jth element the common offset for the jth attribute. If it is possible to estimate the offsets directly, the optimal model can be found by subtracting the offsets from the configurations and use the bilinear model of the average corrected configuration. The offsets, however, are not in general known, and hence have to be handled in modeling the data [10]. The above model assumes that the offset for a given attribute is the same for all panelists. If every panelist has his or her own individual set of offsets, then the structural model will be

$$\mathbf{X}_k = \mathbf{TP}^{\mathrm{T}} + \mathbf{1}\mathbf{m}_k^{\mathrm{T}} + \mathbf{E}_k, \quad k = 1, \ldots, K \qquad (6)$$

where **1** is an I-vector of ones and $\mathbf{m}_k$ is J-vector holding in its jth element the offset of the kth panelist for the jth attribute. This model may seem to be the most plausible for describing offsets in sensory data, since panelists are unlikely to have identical offsets. However, all mathematical and statistical models of fallible data are based on finding the best compromise between fit and variance. Even though the existence of individual offsets is a plausible model from a sensory point of view, it may not always make sense from a modeling point of view. If the offsets of the panelists are very alike (and/or near zero) and only a few samples are assessed, then the uncertainty associated with estimating the offsets individually may in practice lead to overfitting. In such a situation it may happen that the bias introduced by assuming the same offsets for each panelist is smaller than the error introduced by the variance in estimating the offsets individually.

The two approaches for handling offsets (individual or common) are nested in the sense that the model with individual offsets also covers the common offsets. Suppose that the model

$$\mathbf{X}_k = \mathbf{TP}^{\mathrm{T}} + \mathbf{1m}_k^{\mathrm{T}} + \mathbf{1m}^{\mathrm{T}} + \mathbf{E} \quad k = 1, \ldots, K \tag{7}$$

is sought. Setting $\mathbf{n}_k = \mathbf{m}_k + \mathbf{m}$, it is easy to realize that the common offsets are already implicitly given in the individual offsets. Therefore, one only needs to consider one of the two alternatives not the combination. In sensory practice, the model with individual offsets is mostly assumed and indirectly fitted by first centering each configuration matrix across samples. That is, the offsets are handled by removing the corresponding means and the remaining parameters are fitted to the residuals upon centering. This can be shown to provide an overall least squares model in the situations encountered here when individual offsets are present for each panelist [10,11].

2.2. Scenario B: the panelists use different scales

When the panelists differ only in having different overall sensitivity (or sign switches or variability) this can be accommodated by adjusting the ideal model structure as

$$\mathbf{X}_k = \mathrm{a}_k\mathbf{TP}^{\mathrm{T}} + \mathbf{E} \quad k = 1, \ldots, K \tag{8}$$

i.e. every panelist behaves according to the exact same model except that their magnitudes of scoring differ. This scaling factor is similar to the isotropic scaling factor often used in generalized Procrustes analysis [12]. Note that a scaling difference and variability difference has the same influence on how the underlying scores are found because both can be handled by appropriately weighting each configuration matrix.

When the panelists agree on scoring to the extent that they use the same latent variables, only they have different sensitivity towards these *latent* variables, then the model becomes

$$\mathbf{X}_k = \mathbf{TD}_k\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \quad k = 1, \ldots, K \tag{9}$$

where $\mathbf{D}_k$ is a diagonal matrix, the *f*th diagonal elements holding the weight for the *f*th latent variable of the *k*th panelist.

This model introduces different so-called saliencies for each panelist with respect to each latent variable, while the loadings defining the latent variables are the same for all panelists. Specifically, whereas according to this model all panelists use the same set of latent variables, the extent to which they use each particular latent variable to distinguish between products may differ. For instance, one panelist may mainly distinguish products as to their bitterness and hardly distinguish them as to their texture, whereas another may mainly distinguish products in terms of their texture, and hardly distinguish them as to their bitterness On the manifest level, the influence of different saliencies of different latent variables will not be discernable straightforwardly. However, on the latent level the differences in sensitivity are easily described. This model is equivalent to the PARAFAC/CANDECOMP model also underlying the INDSCAL model. PARAFAC has been used extensively in chemometrics [13–16]. PARAFAC was suggested earlier [1] as one possible multi-way extension of PCA in sensory data analysis, though no results were shown. Reference [17] used PARAFAC for modeling data from sensory evaluation of Aceto Balsamico Tradizionale di Modena (ABTM) and also compared with alternative PCA models. It was found that PARAFAC offered more detailed insight on the data. No specific rationale was given for the choice of model in the paper, except that PARAFAC was chosen due to the three-way nature of the data.

Note that high variance does not imply high salience. The structure of the PARAFAC model implies that only variation that is consistent with the *latent* variables lead to high saliences. An assessor who uses the scale to a high degree but fails to be in accordance with the remaining assessors will have low saliencies.

Like for the offsets, the two above models are hierarchically related because the former can be written as $a_k\mathbf{TP}^{\mathrm{T}} = \mathbf{T}\,a_k\mathbf{IP}^{\mathrm{T}}$ with $\mathbf{I}$ being an identity matrix. Hence the former model is a constrained version of the latter. Thus, the isotropic scaling is already implicitly given in the saliencies and hence it is not necessary to incorporate it into the model (Figure 1).

2.3. Combining scenarios for specific models

For a given problem, a suitable model may consist of elements from both scenarios. It is possible to combine the different parts

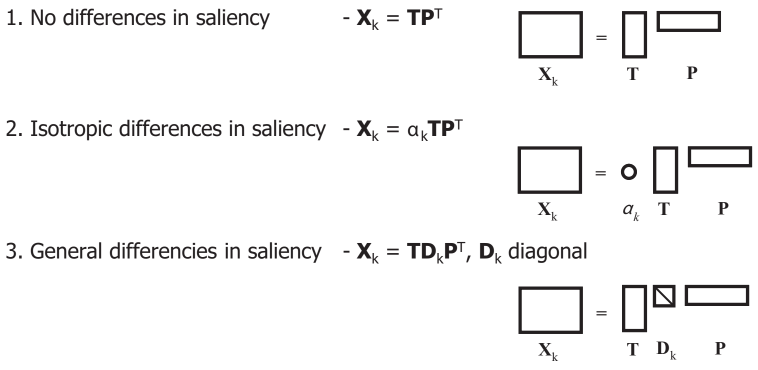
1. No differences in saliency - $\mathbf{X}_k = \mathbf{TP}^{\mathrm{T}}$
$\mathbf{X}_k$
=
T
P
2. Isotropic differences in saliency - $\mathbf{X}_k = \alpha_k\mathbf{TP}^{\mathrm{T}}$
$\mathbf{X}_k$
=
a_k
T
P
3. General differencies in saliency - $\mathbf{X}_k = \mathbf{TD}_k\mathbf{P}^{\mathrm{T}}$, $\mathbf{D}_k$ diagonal
$\mathbf{X}_k$
=
T
$\mathbf{D}_k$
P

Figure 1. Graphical illustration of different sensitivities.

into one model. Each such combination gives rise to a new model. In order to establish the most suitable model, it is necessary to select (i) type of offset, (ii) type of salience modeling and (iii) number of components. It is not suggested that, in practice, all these combinations are investigated. However, some common sense based on empirical results may be obtained, to guide in choosing appropriate models. Quite likely, good starting points for an analysis would be an ordinary PCA model on average centered data and a similar PARAFAC model. Based on the results of two such appropriately validated models, it is usually possible to get indications as to whether these are suitable or whether more alternatives should be investigated. If these two 'extremes' provide similar result it is unlikely that significantly different results can be obtained from intermediate complexity models [8].

Several approaches to choosing the proper model complexity and validating the model are possible. Both for PCA and PARAFAC many diagnostics have been developed and these will not be repeated here but merely used where appropriate [13,18–26].

3. EXPERIMENTAL

3.1. Data analysis

All analyses were performed in MATLAB (Mathworks, Inc.) using the *N*-way Toolbox [27] as well as additional in-house made routines. All data and algorithms are available at www.models.life.ku.dk (Feb. 2007). Two competing models were tested here. The first one, PCA, was performed by taking the average (over panelists) of the data matrices centered across samples and then scaling each attribute variable to unit standard deviation. The second one, PARAFAC, was performed on the three-way data, first centered across samples and then scaled within the attributes to have unit sum of squares per attribute. The averaging in the two cases is identical, whereas the scaling differs marginally. It is

Table I. Cream cheeses investigated in the study. Name and abbreviations applied in plots and text

Products	Abbreviation
Standard full fat cream cheese	'34%'
Medium fat reduced cream cheese	'24%'
Maximum fat reduced cream cheese	'16%'
Prototype cream cheese	P
Prototype cream cheese + Butter Aroma	P + Aroma
Commercial cream cheese A	A + Prot
Commercial cream cheese B	B + Prot
Commercial cream cheese C	C + CHO
Commercial cream cheese D	D + CHO

Table II. Sensory descriptors (with the original Danish word) used in descriptive analysis, abbreviations applied in plots, definitions and reference materials (if provided)

Descriptors (original Danish words)	Abbreviation in plots	Definition and reference material where provided
By olfaction (Nose)		
Cream aroma (fløde lugt)	N-Cream	Reference: 38% fat cream
Acidic aroma (syrlig lugt)	N-Acidic	Reference: 3.5% fat plain yoghurt
Butter aroma (smør lugt)	N-Butter	Reference: Butter
Old milk aroma (gammel mælk lugt)	N-Old milk	Reference: 3.5% fat milk gone sour
By vision (EYES)		
White (hvid)	E-White	The color white
Grey (grå)	E-Grey	The color gray
Yellow (gul)	E-Yellow	The color yellow
Green (grøn)	E-Green	The color green
Grainy (grynet)	E-Grainy	Coarseness after spreading
Glossy (blank)	E-Glossy	Glossiness after spreading
By HAND		
Resistance (modstand)	H-Resistance	Resistance during spreading with knife
By MOUTH*		
Firm (fast)	M-Firm	Hardness of sample in first press with tongue against palate
Melt down (nedsmeltning)	M-Melt down	Rate of melting of product
Resistance (modstand)	M-Resistance	Used force to dissolve food bolus
Creaminess (cremet)	M-Creaminess	Creaminess sensation in mouth
Grainy (grynet)	M-Grainy	Grainy sensation in mouth
Chalky (kridtet)	M-Chalky	Chalky sensation (type astringency)
Cream flavor (flødesmag)	M-Cream	Cream flavor
Fat flavor (fed smag)	M-Fat	Fat flavor
Butter flavor (smør smag)	M-Butter	Butter flavor
Salt (salt)	M-Salt	Salt taste
Sour (sur)	M-Sour	Sour taste
Sweet (sød)	M-Sweet	Sweet taste

*This encompasses both flavor, texture and taste modalities.

Table III. Sensory descriptors, mean values (over panelists and replicates, $n = 24$) for products. Tukey's Honestly Significant Differences values ($p < 0.05$), indicating least significant difference between products in individual descriptors

Descriptors	*F*-value (df = 9,63)	*p*-value	Tukey's HSD (5%)	Products									
				16%	24%	33%	P	P + Aroma	A + Prot 1	A + Prot 2	B + Prot	C + CHO	D + CHO
N-Cream	1.53	0.158NS	1.23	7.40	7.39	7.75	7.40	8.19	7.26	7.24	7.10	7.95	8.09
N-Acidic	1.78	0.072NS	1.28	7.95	8.16	8.05	7.99	7.84	8.85	8.84	7.79	8.45	7.98
N-Butter	6.14	***	1.35	8.17	8.29	8.42	8.99	9.31	8.12	8.92	7.93	9.42	9.53
N-Old milk	2.39	0.021*	0.78	2.93	2.85	2.74	2.46	2.77	3.21	3.20	3.20	2.51	2.44
E-White	17.23	***	0.92	9.78	9.42	7.47	8.73	8.63	10.00	10.14	10.21	7.86	10.41
E-Grey	4.03	***	0.79	2.71	3.16	2.37	2.51	2.57	3.31	3.03	2.89	2.24	2.64
E-Yellow	28.88	***	0.85	3.19	3.74	5.87	4.34	4.06	2.63	2.59	2.51	5.11	2.84
E-Green	1.06	0.402NS	0.41	1.62	1.65	1.84	1.69	1.91	1.72	1.70	1.62	1.62	1.67
H-Resistance	37.98	***	0.92	4.72	8.84	8.78	5.34	5.64	7.57	6.98	6.05	4.71	9.38
E-Grainy	2.40	0.021*	1.14	3.51	4.96	4.81	3.93	3.93	4.96	5.29	4.44	3.70	3.43
E-Glossy	37.36	***	0.97	10.69	8.18	4.84	10.11	9.61	7.14	7.41	8.47	9.29	5.36
M-Firm	43.10	***	0.99	5.12	9.21	8.32	5.54	5.28	8.64	8.85	6.32	4.18	10.11
M-Melting	26.97	***	1.32	9.13	5.96	7.71	9.23	9.52	6.68	6.06	7.88	9.66	5.34
M-Resistance	20.63	***	1.31	5.11	8.19	6.46	4.89	4.66	7.47	8.31	6.38	4.31	9.18
M-Creaminess	3.03	0.005**	1.17	7.86	8.26	8.30	7.12	7.31	6.80	6.91	7.78	8.46	9.09
M-Grainy	4.23	***	0.97	3.66	3.63	3.69	3.58	3.40	3.09	2.71	2.25	2.32	2.20
M-Chalky	32.62	***	1.12	3.95	4.16	4.03	3.86	4.08	7.84	7.67	5.29	3.09	3.46
M-Cream	5.35	***	0.98	7.88	7.85	8.42	7.83	7.99	6.49	7.04	6.38	7.76	7.97
M-Fat	8.15	***	1.11	8.24	8.88	8.86	8.73	8.97	7.01	7.26	6.83	8.99	9.56
M-Butter	17.30	***	0.96	8.06	8.51	9.28	9.05	9.29	6.49	6.78	5.64	9.44	8.87
M-Salt	10.37	***	0.89	6.54	6.30	6.76	6.46	6.30	5.27	5.31	5.26	8.19	5.39
M-Sour	8.99	***	1.09	7.64	7.35	6.73	7.33	7.36	7.97	8.13	9.70	6.64	6.93
M-Sweet	2.59	0.013*	0.73	3.56	3.25	3.93	3.24	3.38	2.92	2.71	3.08	3.21	3.49

customary to use root squares rather than standard deviations in three-way preprocessing [10,28] for different numerical reasons. In this case, scaling to unity of one or the other, though, gave identical results.

3.2. Data

Data were obtained on an earlier occasion [29]. For convenience and aid in plot reading, an overview of the data is provided here. Table I lists nine cream cheeses evaluated in the study. The samples were a mixture of commercially available and specifically produced samples. The main objectives were to assess (1) how the sensory properties are affected by changes in fat level; (2) differences between two main types of fat mimetics protein-based and carbohydrate-based and (3) whether addition of a cream aroma would affect other sensory properties than just aroma/flavor descriptors.

3.3. Sensory descriptive analysis methodology

Sensory evaluation was performed in three replicates. All sessions took place in the sensory laboratory at Department of Food Science, Copenhagen University (KU), which complies with ISO standards [30]. A panel consisting of eight external paid panelists was used for the evaluation. All panelists were part of KU's sensory laboratory standing panel and are tested, selected and trained according to ISO standards [31]. In five sessions panelists were trained on the products, and descriptors were chosen after suggestions from the panel leader on the basis of consensus among the panelists. In the fifth training session panelists evaluated a subset of the samples for sensory evaluation in the sensory evaluation booths. Twenty-three descriptors were developed for the descriptive analysis. Table II lists descriptors in English, their original Danish words and reference materials if provided. For a detailed description of procedures, please confer Reference [29]. For all evaluation sessions a computerized score collection software (FIZZ version 1.30, Biosystemes, France) was used. For all descriptors a horizontal 15 cm unstructured line scale was used. Nine products were tested, with five products in a session, so two sessions equaled one full replicate. This set-up allowed for the use of an internal reference to test the stability of panel evaluations in different sessions. Product A+ Prot was chosen as hidden internal reference and this product was evaluated in all six sessions. Sensory analysis of the 10 (eight different products and two of the same product) products was carried out in triplicate and in randomized order over panelists within each replicate (Latin square design).

3.4. ANOVA

For initial analysis and for illustration, a descriptive univariate analysis was performed. ANOVAs for the individual descriptors were performed with a mixed model, with products as fixed effect and panelists as random factor. Tukey's Honestly Significant Difference test (HSD) for multiple comparison tests was performed on descriptors with significant product differences ($p < 0.05$). Tukey's HSD assures against making claims of any statistically significant result when it may just be the result of chance variation, following that multiple comparisons are made. The size of Tukey's HSD is thus larger than Fisher's LSD. The latter is the least conservative estimate of *post hoc* differences. Univariate analyses were performed in SPSS (ver. 10.0.0, SPSS, Inc. Chicago, IL., USA). Note, however, that such analysis has a different scope than what is the scope here. With multivariate models relations between descriptors are sought whereas in traditional ANOVA such are explicitly ignored. Such multivariate and univariate models are to be considered complimentary rather than competitive and in this paper we are specifically interested in how to model latent underlying structure in the data.

Univariate analysis of the sensory data indicated that 20 of the 23 descriptors varied significantly over the nine products. The descriptors N-Cream, N-Acidic and E-Green were non-significant. Table III lists means and Tukey's Honestly Significant Difference at 5% level for all samples over all panelists and replicates. It shows differences in all sensory modalities (appearance, olfaction, tactile, texture and taste) between products.

4. RESULTS

4.1. PCA results

A PCA model with two components was deemed appropriate for describing the preprocessed and averaged data (Figure 2). Note in the figure that there is a huge difference in the percentages depending on whether these are given in terms of the averaged data or in terms of the individual configurations. The two set of bars in the plot show how well the PCA model of the averaged data fit the same data and how well it fits the original configurations. When reported in terms of averaged data, the percentages do not say anything about how well the panelists' individual differences are modeled as these are purposely 'neglected' by averaging. Thus, the much higher apparent explained variation is simply because it only describes the fit to panelist-averages.

In Figure 3, the scores and loadings from the PCA model are shown. As can be seen replicates are close for some of the

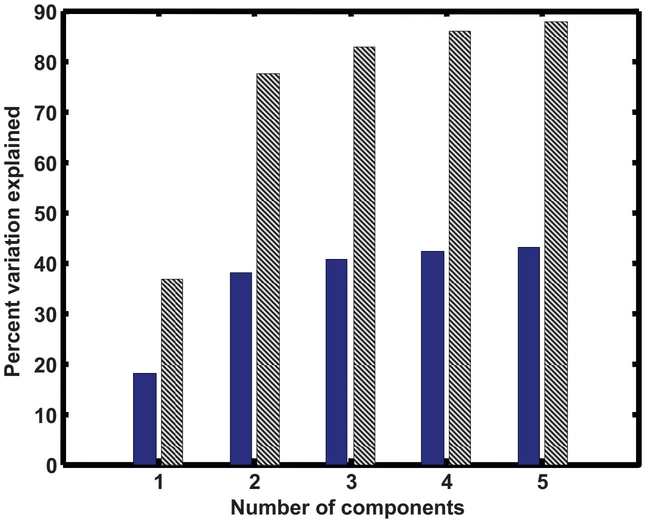

Figure 2. Percentage variation explained by PCA in terms of averaged data (right-most bars) and in terms of the non-averaged data (left-most bars). This figure is available in colour online at www.interscience.wiley.com/journal/cem

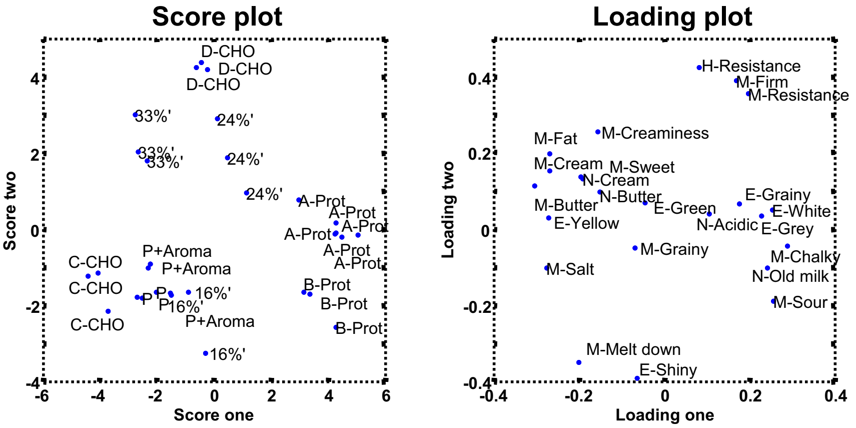

Figure 3. Scores and loadings from a two-component PCA model. This figure is available in colour online at www.interscience.wiley.com/journal/cem

products (e.g. D-CHO), while others are more dispersed (e.g. 24%). This indicates that some of the products are evaluated more similarly between replicates. The underlying reason is probably higher intrinsic variation for some products, but it is confounded with the session uncertainty. The relatively low dispersion of product A-Prot (hidden internal standard), supports that it is higher product variation that leads to the variations between replicates. The loading plot shows that the first component overall relates to differences in sour taste (M-Sour) and chalky sensation (M-Chalky, a type of mouth feel strongly related to astringency), and a group of Fat-related flavor properties (M-Butter, M-Cream, M-Fat and M-Creaminess), and that these differences also manifest themselves in the appearance of the products (E-White and E-Grey vs. E-Yellow). The second component relates to differences in texture properties. Products with a firm texture (M-Firm and H-Resistance), that resists break down in the mouth (M-Resistance) are in the upper part of the product space, while soft products with a faster melt down (M-Melt down) are in the opposing end of the sensory property space.

4.2. PARAFAC results

PARAFAC models with increasing numbers of components were fit to the preprocessed data. Standard validation routines such as core consistency [18], residual analysis [25,32] etc. were employed but will not be described in detail here. For comparison only, a plot similar to Figure 2 is shown in Figure 4. It is noted that

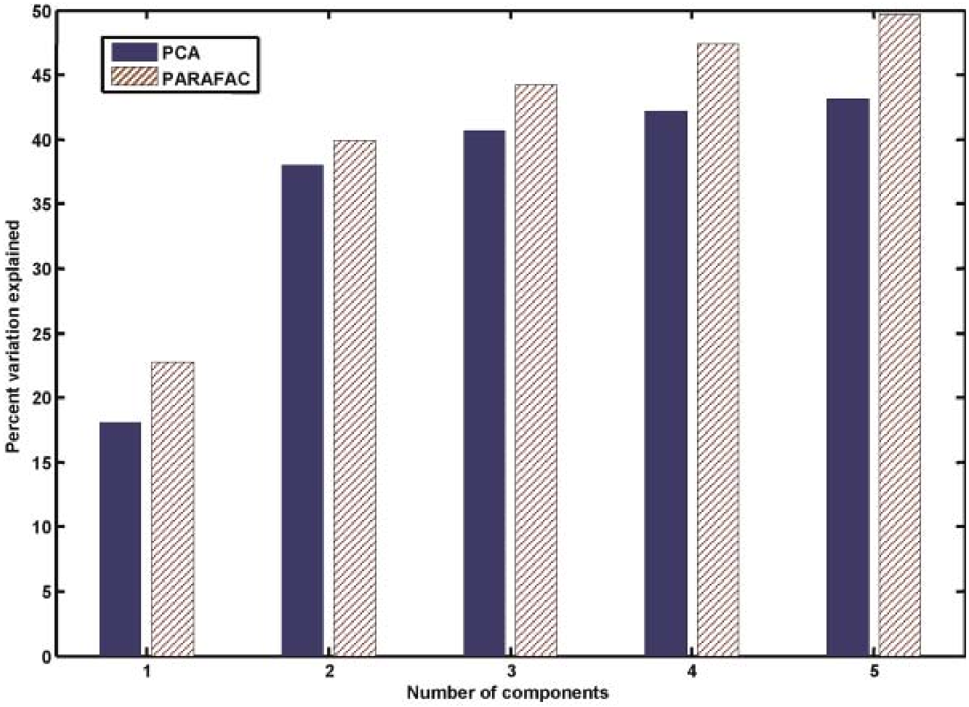

Figure 4. Percentage variation explained by PCA (left-most bars) and PARAFAC (right-most bars). This figure is available in colour online at www.interscience.wiley.com/journal/cem

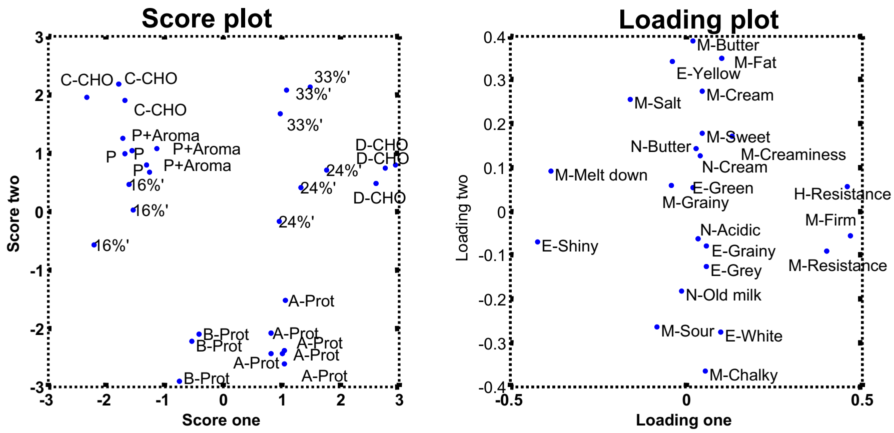

Figure 5. PARAFAC product score and attribute loading one versus two. This figure is available in colour online at www.interscience.wiley.com/journal/cem

PARAFAC fits the data better than PCA, but this is a trivial consequence of the PARAFAC model including more parameters per component (the saliences). The number of components was chosen to be two, although subtle improvements in fit were possible with a third component. For the sake of simplicity and to focus on the main variation though, only two components are used in the following.

Looking at the PARAFAC results (Figure 5), a rotation compared to PCA is observed. This is mainly due to the fact that PCA results will have an 'arbitrary' rotation determined by the additional mathematical constraints of PCA (maximum variance per component and orthogonality). Such restrictions are not needed nor used in PARAFAC which is why the specific rotation of axes in the two models is not expected to be similar. Disregarding the rotation, a similar product and descriptor pattern is observed (Figure 5) indicating that the two models are modeling similar variation and hence also indicating that there is no need to look further into other alternative models.

Additional information about the panelists can be seen in the panelist loadings/saliencies (Figure 6). Along the first component the panelists are dispersed in a continuum, but along the second component the panelists form two distinct groups.

This corresponds to the fact that the response to some texture descriptors (M-Melt down, H-Resistance, M-Firm and M-Resistance) is generally agreed upon, but the panelists respond to them with more or less emphasis, i.e. *panelists with highest loadings seen for these four descriptors use a larger part of the scale for their evaluation*. Contrary to that, the panelists are grouped in two when it comes to the descriptors in the latent variable represented by loading two. This is most clearly expressed in M-Chalky and the group of fat-related flavor descriptors (M-Butter, M-Fat and M-Cream). The panelist loadings indicate that some panelists are more consistent than others. Indeed this can be verified in the raw data. In Figure 7 it is easily seen that the low-loading panelists 5 and 8 are not consistent on the descriptor M-Butter which is one of the important descriptors for the second component. Panelist 1 and 2, on the other hand, are much more consistent and use a larger part of the scale. However, the differences could be much larger. For both first and second components, the loadings span from approximately 0.22 to 0.45. Had the panelists completely disagreed on the use and meaning of the descriptors i.e. some panelists reversed the order of products, the span would have been larger and also included negative loadings. We have observed this in other sensory profiles, when analyzing individual differences in perception of creaminess.

The loading of a specific panelist on a specific component is a measure of how descriptive or consistent the panelist is in judging the latent variable under consideration. Thus, an overall measure of the quality in terms of saliency of a panelist can be obtained by looking at the magnitude across all latent variables. In Figure 8 the sum of squares of loadings per panelist is shown relative to the one with largest sum of squares. As can be seen, panelist 4 and 7 are assessed as most consistent whereas, e.g. 3, 5 and 8 are the least consistent ones (see Figure 6). Judge 2 and 6 are equally consistent but on opposite components and obtain the same overall quality. The quality measure reflects the consistency of each assessor with respect to the overall model and hence assumes that the model is valid.

5. CONCLUSION

A new set of models has been suggested. Most notably, the PARAFAC model is suggested to have an intuitive interpretation

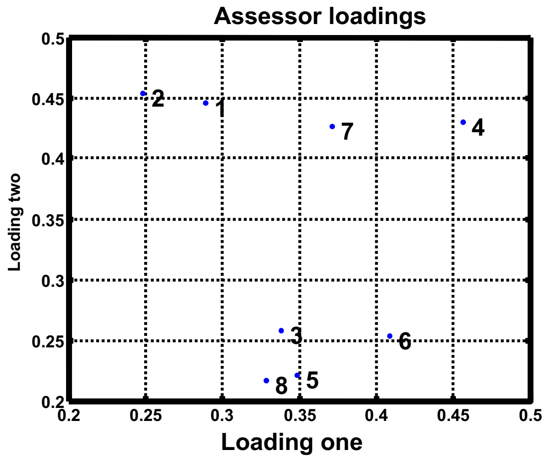

Figure 6. PARAFAC panelist loading one versus two. Each point is located at the coordinates from the diagonal of $\mathbf{D}_k$, of the corresponding assessor. This figure is available in colour online at www.interscience.wiley.com/journal/cem

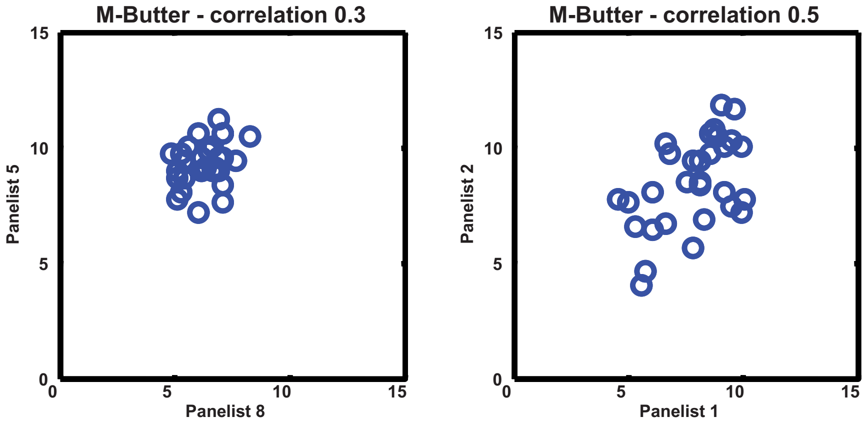

Figure 7. Evaluations of two panelists who are performing poorly (left) and two panelists who are performing well (right) for the descriptor M-Butter. This figure is available in colour online at www.interscience.wiley.com/journal/cem

for sensory data by introducing assessor-specific parameters. In the example, it was shown that PCA and PARAFAC gave quite similar interpretations with respect to samples and descriptors. This can be expected for data from a well-trained panel evaluating a set of samples with large sensory differences. However, the PCA analysis is explicitly based on the assumption that all panelists are equally good, i.e. that they agree and do not exhibit significant individual differences. The PARAFAC model on the other hand assumes that different assessors have different sensitivities towards different latent variables and this enables an intuitive and meaningful handling of variations in scale and variability between assessors. Together with appropriate centering, it is then possible to handle most of the expected individual differences in latent variable modeling of sensory data. The PARAFAC model (i) estimates the saliencies, (ii) provide sample and descriptor loadings taking these differences into account and (iii) provides tools for visualizing and assessing the differences.

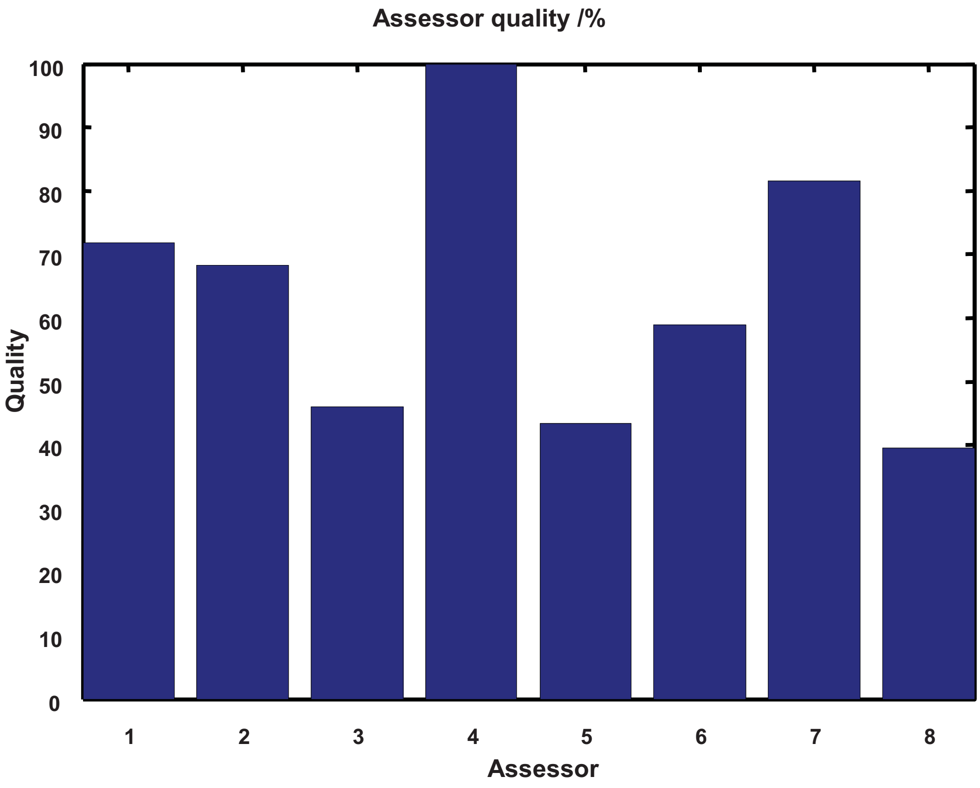

Figure 8. Panelist quality judged by relative size of loadings. This figure is available in colour online at www.interscience.wiley.com/journal/cem

REFERENCES

1. Brockhoff PMB, Hirst D, Næs T. Analysing individual profiles by three-way factor analysis. In Multivariate Analysis of Data in Sensory Science, Næs T, Risvik E (eds). Elsevier: Amsterdam, 1996; 307–342.
2. Brockhoff PMB, Skovgaard I. Modelling individual differences between assessors in sensory evaluations. *Food Qual. Prefer.* 1994; **5**: 215–224.
3. Næs T. Handling individual differences between assessors in sensory profiling. *Food Qual. Prefer.* 1990; **2**: 187–199.
4. Næs T, Hirst D, Baardseth P. Using cumulative ranks to detect individual differences in sensory profiling. *J. Sens. Stud.* 1994; **9**: 87–99.
5. Brockhoff PMB. Statistical testing of individual differences in sensory profiling. *Food Qual. Prefer.* 2003; **14**: 425–434.
6. Carroll JD, Chang J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika* 1970; **35**: 283–319.
7. ten Berge JMF, Kiers HAL. Some clarification of the CANDECOMP algorithm applied to INDSCAL. *Psychometrika* 1991; **56**: 317–326.
8. Kiers HAL. Hierarchical relations among three-way methods. *Psychometrika* 1991; **56**: 449–470.
9. Qannari EM, Wakeling I, MacFie HJH. A hierarchy of models for analysing sensory data. *Food Qual. Prefer.* 1995; **6**: 309–314.
10. Bro R, Smilde AK. Centering and scaling in component analysis. *J. Chemom.* 2003; **17**: 16–33.
11. Kruskal JB. *Some least squares theorems for matrices and N-way arrays* Manuscript, Bell Laboratories, Murray Hill, New Jersey 1977.
12. Gower JC. Generalized procrustes analysis. *Psychometrika* 1975; **40**: 33–51.
13. Bro R, PARAFAC Tutorial and applications. *Chemom. Intell. Lab. Syst.* 1997; **38**: 149–171.
14. da Silva JCGE, Novais SA. Trilinear PARAFAC decomposition of synchronous fluorescence spectra of mixtures of the major metabolites of acetylsalicylic acid. *Analyst* 1998; **123**: 2067–2070.
15. Durell SR, Lee C, Ross RT, Gross EL. Factor analysis of the near-ultraviolet absorption spectrum of plastocyanin using bilinear, trilinear, and quadrilinear models. *Arch. biochem. biophys.* 1990; **278**: 148–160.
16. Sanchez E, Kowalski BR. Tensorial resolution: a direct trilinear decomposition. *J. Chemom.* 1990; **4**: 29–45.
17. Cocchi M, Bro R, Durante C, Manzini D, Marchetti A, Saccani F, Sighinolfi S, Ulrici A. Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. *Food Qual. Prefer.* 2006; **17**: 419–428.
18. Bro R, Kiers HAL. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* 2003; **17**: 274–286.
19. Eastment HT, Krzanowski WJ. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* 1982; **24**: 73–77.
20. Harshman RA, De Sarbo WS. An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In Research Methods for Multimode Data Analysis, Law HG, Snyder CW, Hattie JA, McDonald RP (eds). Praeger Special Studies: New York, 1984; 602–642.
21. Jackson J. A User's Guide to Principal Components. Wiley & Sons: New York, 1991.
22. Louwerse DJ, Kiers HAL, Smilde AK. Cross-validation of multi-way component models. *J. Chemom.* 1999; **13**: 491–510.
23. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. Academic Press: London, 1979.
24. Riu J, Bro R. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemom. Intell. Lab. Syst.* 2003; **65**: 35–49.
25. Smilde AK, Bro R, Geladi P. Multi-way analysis. Applications in the chemical sciences. Wiley PL New York 2004.
26. Wold S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978; **20**: 397–405.
27. Andersson CA, Bro R. The N-way Toolbox for MATLAB. *Chemom. Intell. Lab. Syst.* 2000; **52**: 1–4.
28. Harshman RA, Lundy ME. Data preprocessing and the extended PARAFAC model. In Research Methods for Multimode Data Analysis, Law HG, Snyder CW Jr, Hattie J, McDonald RP (eds). Praeger: New York, 1984; 216–284.
29. Frøst MB. *The influence of fat content on sensory properties and consumer perception of dairy products*. Ph.D. thesis, The Royal Veterinary and Agricultural University, Centre for Advanced Food Studies, Department of Dairy and Food Science, 2002; 1–157.
30. ISO-8589, International Standard 8589. *Sensory Analysis–General guidance for the design of test rooms* International Organisation for Standardisation (Genéve) 1988.
31. ISO-8586-1, International Standard 8586-1. *Sensory analysis–Methodology–general guidance for the selection, training and monitoring of assessors* International Organisation for standardisation, Genève 1993.
32. Harshman RA. How can I know it's real? A catalog of diagnostics for use with three-mode factor analysis and multidimensional scaling. In Research Methods for Multimode Data Analysis, Law HG, Snyder CW Jr, Hattie J, McDonald RP (eds). Praeger: New York, 1984; 566–591.