Food Quality and Preference 20 (2009) 167–175

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Interpreting sensory data by combining principal component analysis and analysis of variance

Giorgio Luciano [a], Tormod Næs [a,b,*]

[a] NOFIMA FOOD, Matforsk, Oslovegen 1, 1430 Ås, Norway
[b] Department of Mathematics, University of Oslo, Blindern, Oslo, Norway

ARTICLE INFO

Article history:
Received 13 June 2008
Received in revised form 15 August 2008
Accepted 19 August 2008
Available online 11 September 2008

Keywords:
PCA
ANOVA
Sensory profiling
ASCA

ABSTRACT

This paper compares two different methods for combining PCA and ANOVA for sensory profiling data. One of the methods is based on first using PCA on raw data and then relating dominating principal components to the design variables. The other method is based on first estimating ANOVA effects and then using PCA to analyse the different effect matrices. The properties of the methods are discussed and they are compared on a data set based on sensory analysis of a candy product. Some new plots are also proposed for improved interpretation of results.

1. Introduction

Sensory panel data can always be looked upon as three-way data tables with assessors, objects/samples and attributes as the three "ways". In order to analyse differences and similarities between samples and assessors as well as the correlation structure among attributes, the three-way structure of the data needs to be taken into account. This can be done in various ways using different underlying ideas and philosophies.

A technique that can be useful in some cases is regular multivariate analysis of variance (MANOVA, Kent & Bibby, 1978) for testing the effect of samples and/or assessors for all attributes simultaneously. Usually one is, however, interested in more insight than this method can give and therefore other techniques are to be preferred. A much used method within the area of sensory analysis is the generalised procrustes analysis (GPA), treating each assessor slice as a matrix, followed by a principal component analysis of the average or consensus matrix (Dijksterhuis, 1996). GPA is based on the idea of making individual assessor data matrices as similar as possible to each other by scaling and rotation. Another possible approach is regular principal components analysis (PCA) of all individual sensory profiles followed by a two-way ANOVA of the most important components with assessor and products effects as independent variables. (Ellekjær, Ilseng, & Næs, 2002). The rows in the data table used for this analysis correspond to all samples * assessor combinations and the columns correspond to sensory attributes. This method can be modified using the 50–50 MANOVA (Langsrud, 2002) method which handles significance testing in a more elegant way. Using partial least squares regression (PLS-2) of all sensory profiles versus the two independent design variables assessors and products and their interaction is a closely related approach (Martens & Martens, 2001). PCA based on the alternative unfolding with objects and assessor * attributes as columns and rows has been tested in for instance Dahl and Næs (2006). In the same paper a generalised canonical correlation analysis CCA (Carroll, 1968) analysis of individual sensory data was tested and compared to PCA. Classical three-way factor analyses such as Tucker-2 and PARAFAC have also found useful applications within the framework of sensory analysis (Bro, Qannari, Kiers, Næs, & Frøst, 2008; Brockhoff, Hirst, & Næs, 1996). Recently an alternative method for three-way analysis of variance (ANOVA) has been proposed in the chemometric literature (ASCA, Jansen et al., 2005), but the method is not yet tested for sensory data. ASCA is a method which first uses regular two-way ANOVA for each attribute separately, estimates the effects (under regular ANOVA restrictions) and then uses PCA on the main effects matrices and interaction matrix separately for interpretation of results. The method has recently been combined with PARAFAC in the so-called PARAFASCA (Jansen et al., 2008). Other important approaches and overviews of alternative methods can be found in Qannari, Wakeling, Courcoux, and MacFie (2000, 2001) and in Hanafi and Kiers (2006).

The present paper is a comparison study of two of the ANOVA based methods described above. In particular we will be interested in comparing the newly developed ASCA method with traditional PCA of the unfolded three-way data table followed by ANOVA (here

* Corresponding author. Tel.: +47 64 97 0333; fax: +47 64 97 0165.
E-mail address: tormod.naes@matforsk.no (T. Næs).

0950-3293/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodqual.2008.08.003

called PC-ANOVA). As can be noted, the two methods are closely related in the sense that they are both based on the same two basic methodologies, two-way ANOVA and PCA, but with the difference that the two methodologies are used in opposite order. These approaches have the advantage over other methods that they focus both on the multivariate aspects of the sensory profiles and the explicit relation of the sensory data to the design of the study. The methods will be compared conceptually and also with respect to results obtained in an empirical illustration.

2. Theory

In the present paper we will consider a three-way data table with $I^* M$ rows corresponding to M replicates of I samples, K columns corresponding to the attributes and with J slices corresponding to assessors. We refer to Fig. 1 for an illustration of the structure of the data set.

Three way data of this type can always, for each attribute k, be modelled by the two-way ANOVA model

$$X^k_{ijm} = \mu^k + \alpha^k_i + \beta^k_j + \alpha\beta^k_{ij} + e^k_{ijm} \tag{1}$$

Here the μ is the general mean, the α's are main effects for products, the β's main effects for assessors, the $\alpha\beta$'s the interactions and e is the random error term corresponding to replicate variation. For ANOVA purposes, the error terms are assumed to be uncorrelated and normally distributed with the same variance. The usual way of applying this model is to assume that the assessor and interaction effects are random, leading to a mixed model (see Næs & Langsrud, 1998).

Sometimes experimental designs are used for the samples and in such studies (Baardseth et al., 1992), the product effect can be split in several components corresponding to the experimental factors in the design (Box, Hunder, & Hunter, 1978). How to handle this extension within the framework of the methodologies presented here will be discussed below. How to handle structures in the replicates will also be discussed in the same sections.

In the following we will use the symbol $\mathbf{X}$ to denote the unfolded three-way data table with $I^* M^* J$ rows and K columns. Using this symbol it is possible to rewrite the model in Eq. (1) for all the attributes simultaneously as follows

$$\mathbf{X} = \mathbf{1}\mu^{\mathbf{t}} + \mathbf{D}_1\mathrm{B}_1 + \mathbf{D_2}\mathbf{B}_2 + \mathbf{D}_{12}\mathrm{B}_{12} + \mathbf{E} \tag{2}$$

where μ is the general mean vector for all K attributes simultaneously, 1 is a vector of 1's, the $\mathbf{D_1}$, $\mathbf{D_2}$ and $\mathbf{D_{12}}$ are the dummy design matrices for the products, assessors and interactions between assessors and products respectively and the **B**'s are the corresponding parameter matrices. The $\mathbf{B_1}$ corresponds to the α's in Eq. (1), the $\mathbf{B_2}$ to the β's and $\mathbf{B_3}$ to the $\alpha\beta$'s. The design matrix $\mathbf{D_1}$ will have one column for each assessor and consists of 0's and 1's with a 1 in column j and row i if this line corresponds to an observation for assessor j. The same structure holds for the other two matrices. The matrix **E** is the matrix of residuals. Correlations between the different elements (columns) of this matrix are possible (Mardia, Kent, & Bibby, 1978), but it is always assumed within multivariate ANOVA that the error terms for different observations and replicates are independent.

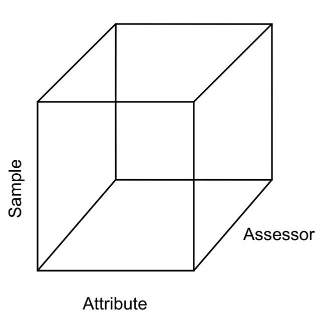

Fig. 1. The data structure for descriptive sensory data.

The model (2) can be used directly, either for each attribute (column) separately or for all simultaneously for testing hypotheses about product and assessor effects. An example of an important hypothesis related to this model is H_0:$\mathbf{B_1} = \mathbf{0}$, which is the hypothesis of no product effects. For the univariate ANOVA, this hypothesis can be separated in K individual hypotheses, one for each attribute. Similar hypotheses can be set up for assessor and interaction effects. If wanted, one can also construct a combined hypothesis of for instance $\mathbf{B_1}$ and $\mathbf{B_{12}}$ as is done in the ASCA paper (Jansen et al., 2005) and in Næs & Langsrud, 1998.

The main problems with regular ANOVA approaches is that they only focus on hypothesis tests and provide little further insight about the relations between the attributes. Therefore ANOVA will usually be accompanied with some type of PCA for further interpretation of the relations between the variables. In this paper we will discuss two alternative approaches proposed in the literature for providing this type of additional insight by combining ANOVA with PCA.

For the purpose of the methods to be discussed below, it is of interest to estimate the effects matrices **B** in Eq. (2) above. This is usually done by least squares (LS) fitting of the responses to the design matrices, but in order to obtain unique results, one needs to add a restriction on the parameter estimates (see e.g. Lea, Næs, & Rødbotten, 1997). This can be done in various ways, but the most common way is to use the restriction that all main effects of assessors and main effects of products sum to 0 and that the same is true for the interactions summed either over assessors or products. In this paper main attention will be given to balanced designs, but how to extend the approach to more general data sets will also be discussed. For the balanced case, the main effects and interactions have a particularly simple expression based on simple averages and subtraction, i.e.

$$\hat{\alpha}^k_i = \bar{X}^k_i - \bar{X}^k \tag{3}$$

$$\hat{\beta}^k_j = \bar{X}^k_j - \bar{X}^k \tag{4}$$

$$\alpha\hat{\beta}^k_{ij} = X^k_{ij} - \bar{X}^k_i - \bar{X}^k_j + \bar{X}^k \tag{5}$$

where $\bar{X}^k_i$ is the average for product i and attribute k, $\bar{X}^k_j$ is the average for assessor j and attribute k. and $\bar{X}^k$ is the total average for attribute k. Note that the interactions can be considered as obtained by double centring of the original data matrix.

When PCA is used in this paper we will always use it on centred data, i.e. data for which the average has been subtracted for each column.

2.1. PCA-ANOVA

The simplest way of combining ANOVA with PCA is to use PCA directly on the unfolded data matrix **X** in Eq. (2) where the number of columns corresponds to the number of attributes and the number of rows corresponds to all assessor, product and replicate combinations. This implies that the PCA gives components that are combinations of all the effects in the model (2). A possibility is to average over replicates before computation of principal components, but generally this is not natural since the replicates are needed for testing purposes in the subsequent ANOVA. Examples of the use of this and similar methodologies can be found in Ellekjær et al. (2002) and in Langsrud (2002).