Food Quality and Preference 42 (2015) 146–153

Contents lists available at ScienceDirect

Food Quality and Preference

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

Check-all-that-apply data analysed by Partial Least Squares regression

Åsmund Rinnan [a,*], Davide Giacalone [b], Michael Bom Frøst [b]

[a] Spectroscopy & Chemometrics Section, Dept. of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg C, Denmark
[b] Sensory & Consumer Science Section, Dept. of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 21 July 2014
Received in revised form 29 January 2015
Accepted 30 January 2015
Available online 7 February 2015

Keywords:
PLS
CATA
Jack-knifing
PLS-DA
Uncertainty
A-PLS

ABSTRACT

This paper discusses the application of Partial Least Squares regression (PLS) to handle sensory data from check-all-that-apply (CATA) questions in a rapid, statistically reliable, and graphically-efficient way. We start by discussing the theory behind the CATA data and how these normally are analysed by multivariate techniques. CATA data can be analysed both by setting the CATA as the **X** and the **Y**. The former is the PLS-Discriminant Analysis (PLS-DA) version, while the latter is the ANOVA-PLS (A-PLS) version. We investigated the difference between these two approaches, concluding that there is none. This is followed by a discussion of how to get a good estimate of the uncertainty of the model parameters in the PLS model. For a PLS model this is often assessed by leave-one-respondent-out cross-validation. We will, though, show that this gives too optimistic uncertainty estimates, and a repeated split-half approach should rather be used. Finally, we will discuss the shortcomings of using univariate techniques such as the Cochran's *Q* test and even the uncertainty estimates based on the Jack-knifed regression coefficients compared to the multivariate reality of the loading weights in PLS-DA. Overall, this paper provides a formal introduction as to how to utilise PLS-DA and cross validation with resampling for the investigation of CATA data.

1. Introduction

1.1. Statistical treatment of data from check-all-that-apply questions

Check-all-that-apply questions (CATA, Adams, Williams, Lancaster, & Foley, 2007) are an increasingly popular technique for fast sensory profiling of food products, which consists in presenting a panelist with a predefined list of attributes, and have them tick all the ones they deem appropriate to describe a given product. CATA questions are used with both trained panels and consumers, but are probably most often associated with latter for their intuitiveness and the little time requirements (Ares, Deliza, Barreiro, Gimenez, & Gámbaro, 2010).

The data produced by this method consist of dichotomous responses (checked attribute = 1; unchecked attribute = 0) for each of the attributes present in the CATA ballot. These responses are arranged in either an unfolded assessor by attribute matrix (Fig. 1A), or a in a cross tabulation matrix containing total frequency of mention for each attribute (Fig. 1B).

Different statistical techniques can be applied to CATA data, depending on whether one is working with the unfolded or the cross tabulation matrix. Cross tabulation matrices are mostly analysed by exploratory multivariate techniques. Correspondence analysis (CA), a factorial method tailored for the analysis of contingency table, is by far the most common analytical tool applied to this type of matrix, and it is used mainly for exploration and graphical visualisation of product differences (e.g., Abdi & Williams, 2010; Clausen, 1998, and Blasius & Greenacre, 2014). The problem with looking at the cross-tabulation matrix is that the model will contain no information with regards to the uncertainty in the different attributes or if the respondents are able to separate the products under investigation. By focusing the analysis on the unfolded matrix these uncertainties comes into play. For sure the explained variance will be lower for the analysis of the unfolded matrix. The reason can be found in the data itself; there is a high degree of uncertainty in the CATA data (especially because this method is commonly applied with untrained respondents). The multivariate data analysis will efficiently separate the information from the noise, but as the noisy part of the data is rather large, this accounts for a larger part of the variation of the data than the informational part. However, and this is important, it is the part of the data holding the information which is of interest.

Because CA is a purely exploratory technique, it is becoming increasingly common to supplement the analysis with univariate tests aimed at detecting significant inter-product differences for

* Corresponding author.
E-mail address: aar@food.ku.dk (Å. Rinnan).

http://dx.doi.org/10.1016/j.foodqual.2015.01.018
0950-3293/

each of the CATA attributes, such as the McNemar's test (Ennis & Ennis, 2013; McNemar, 1947), when comparing only two products, or its extension Cochran's Q test (Cochran, 1950; Patil, 1975), when more than two products are being compared.

Cochran's Q test is a non-parametric procedure used to test whether K treatments have identical effect on a response variable that can take only two possible outcomes (0/1). In relation to CATA data, this corresponds to the set-up shown in Fig. 1A. The Cochran's Q test statistic, following the notation in Fig. 1, is defined by:

Test statistic for the Cochran's Q test.

$$Q_j = \frac{K(K-1)\sum_{k=1}^{K}\left(x_{.k} - \frac{N}{K}\right)^2}{\sum_{i=1}^{I} x_{i.}(K - x_{i.})} \quad (1)$$

where K is the number of products, $x_{.k}$ is the total count for attribute j for the kth product, I is the number of respondents, $x_{i.}$ is the total counts for attribute j for the ith respondent across the K products, and N is the grand total for attribute j. The test consists in checking whether Q_j is larger than the central chi-square criterion for the chosen level of significance with $(K-1)$ degrees of freedom.

On CATA data, Cochran's Q test is usually carried out on an attribute-by-attribute basis, in order to identify significant differences between products (Meyners, Castura, & Carr, 2013). It is important to note, however, that this test does not take into account the inner similarity structure between the different attributes. Also, Cochran's Q test is often followed by pairwise comparisons with related multiplicity issues. Prior to any specific attribute test Meyners et al. (2013) have recently proposed the adoption of an omnibus test, where the test statistic used is the sum of Cochran's Q statistics obtained for individual attributes. They suggest that this number is compared to a manifold randomisation test where the randomisation is performed within each respondent separately; thus maintaining any respondent-to-respondent difference, but investigating if the products themselves were perceived differently by the group of respondents or not. And as such, this test would naturally be followed by the tests mentioned above, but there would then be more certainty that the difference which is searched for in the Cochran's Q test is based on a significant difference between the products.

(A)

Respondent	Product	Attr. 1	Attr. 2	….	Attr. J
1	I	0/1			
1	II				
1	III				
2	I				
2	II				
2	III				
…	…				
I	K				

(B)

Product	Attr. 1	Attr. 2	….	Attr. J
I	Counts			
II				
III				
…				
K				

Fig. 1. Common set-ups for CATA data: unfolded (A) and cross-tabulation (B). I is the number of respondents, K is the number of products and J is the number of attributes.

1.2. Partial Least Squares regression

The goal of this paper is to discuss the use of Partial Least Squares regression (PLS) (Geladi & Kowalski, 1986) for the analysis of category data, such as those CATA delivers. The main goal of CATA is to investigate whether there is a difference between different products in the test set, and in which sensory attributes those differences exist. In addition, it is of interest to understand if there are some attributes which the panel cannot differentiate in the products they have been presented. PLS applied with CATA data can produce intuitive and easy to understand graphical output making it possible for the analyst to assess both quality differences between the products, as well as between the attributes.

1.2.1. PLS and Jack-knifing theory

Partial Least Squares regression (PLS) is a technique which focuses on explaining the variation seen in a response matrix **Y**, from the variation stored in a predictor matrix **X**, with which it shares row dimensionality. This is performed in such a way that the covariance between **X** and **Y** is best explained. However, as this is a regression technique, a model where **X** is used to predict **Y** compared to one where **Y** is used to predict **X** will, at least theoretically, give different results. It is therefore of importance to define the goal of the analysis prior to selecting how the data should be arranged.

Let us define the two matrices first: (1) the design matrix (**Y**) – may include information with regards to product type, respondent and/or consumer groups, and (2) the CATA responses (**X**) – the responses given by the respondents to the different products. There are two options in how to perform an analysis: (1) the design matrix can be used to model the CATA responses (A-PLS) (Martens, Bredie, & Martens, 2000), or (2) the CATA responses can be used to model the design matrix (PLS-DA) (Barker & Rayens, 2003, and Rossini, Verdun, Cariou, Qannari, & Fogliatto, 2012). Model 1 will give information with regards to which sensory attributes discriminates between the products/respondents/consumer groups, while Model 2 will focus on what is relevant for the sensory variation (Martens et al., 2000). These two models are graphically shown in Fig. 2. In this manuscript we will, though, show that the difference from a practical point of view is negligible.

In all mathematical models it is important to validate the model parameters in order not to overfit the model data. Furthermore, validation can be used to get a good estimate of the uncertainty in the model. This uncertainty estimate does not only include looking at the root-mean-squared-error (RMSE) estimate, but could just as easily include uncertainty in the regression coefficients, loadings, scores, etc. The uncertainty can then be used to evaluate the significance of the results achieved in the analysis. Cross-validation (Wold, 1978) is a common method for uncertainty estimates of the prediction error, and Jack-knifing (Martens & Martens, 2000) is performed in much the same way. The only difference between the two is that in normal cross-validation only the calibrated and the validated prediction values are stored, while in Jack-knifing also the individual model parameters on each of the sub-models are stored (i.e., scores, loadings and regression coefficients). The model parameters based on the different sub-models can thus readily be used in order to estimate the uncertainty of these parameters. Martens and Martens (2000) suggested using these model estimates as a variable selection technique where non-significant variables can be removed from the model. In sensory science, this technique has been commonly applied to e.g., descriptive profiling (Martens & Martens, 2001) and time-intensity data (Frøst, Heymann, Bredie, Dijksterhuis, & Martens, 2005).

In this manuscript, we extend this approach to the analysis of CATA data. In particular, we suggest using the Jack-knifed estimates of the scores and loadings in order to evaluate significant