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

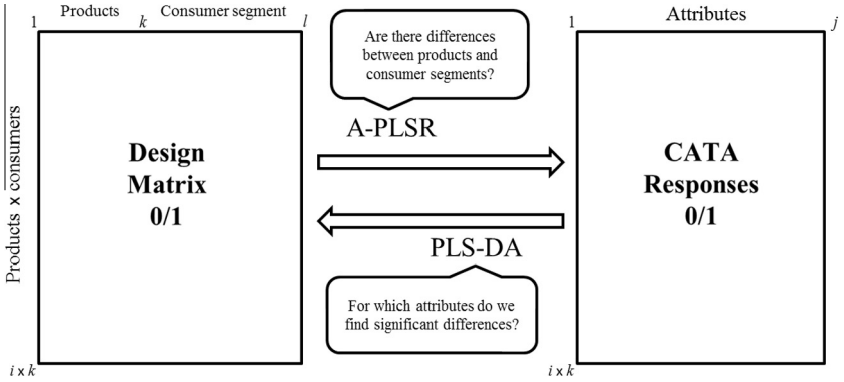

Fig. 2. Data set-up for JK-PLS (adapted from Giacalone, 2013).

differences between products and attributes. However, it is not always necessary to use any resampling method to get a good uncertainty estimate. This is especially the case where the uncertainty in the individual samples (combination of respondents and products in a typical CATA dataset), is smaller than the uncertainty between the products in the original model, i.e., perturbations in the data prior to modelling have little or no impact on the uncertainty estimates for the perceived difference in the products. On the other hand if the uncertainty estimates of interest is for one sample (a specific combination of product and respondent) or variable in the original data-matrix, resampling is of importance. In previous application of PLS on CATA, the chosen resampling strategy has been to use Jack-knifing and define segments corresponding to each respondent (Giacalone, Bredie, & Frøst, 2013; Reinbach, Giacalone, Ribeiro, Bredie, & Frøst, 2014). One potential problem with this strategy is that the uncertainty estimates may get too small as the perturbations to the data are relatively small. There is a correction factor for this inherent in the calculation of the standard error from the Jack-knife estimate (see Eq. (1)), but this correction factor is often not enough. This is a particularly relevant problem for CATA data since this method is commonly applied with a large number of respondents (e.g., 60–80 or more), and thus the difference between each sub-model is very small. A better and more sophisticated method for estimating the uncertainty in the model parameters is bootstrapping (Babamoradi, van den Berg, & Rinnan, 2013; Wehrens, Putter, Lutgarde, & Buydens, 2000). A drawback for this method is that it is not so straightforward to apply. We will therefore suggest an alternative approach where multiple split-half (two segments) Jack-knife runs are used. Jack-knifing is something many practitioners have come across, and as thus is more accepted than the more mathematical bootstrapping. The standard errors (the uncertainty estimate) for each of these methods are given below:

Standard error based on leave-one-respondent-out Jack-knifing.

$$\text{s.e.}_{\text{LORO},\,j} = \sqrt{\frac{A-1}{A} \times \sum_{a=1}^{A} (w_{a,j} - \overline{w_j})^2} \quad (2)$$

Standard error based on bootstrapping.

$$\text{s.e.}_{\text{Boot},\,j} = \sqrt{\frac{1}{A-1} \times \sum_{a=1}^{A} (w_{a,j} - \overline{w_j})^2} \quad (3)$$

Standard error based on repeated split-half Jack-knifing.

$$\text{s.e.}_{\text{RepSH},\,j} = \sqrt{\frac{1}{M} \times \frac{A-1}{A} \times \sum_{a=1}^{A} \sum_{m=1}^{M} (w_{a,m,j} - \overline{w_{.j}})^2} \quad (4)$$

where A is the number of sub models (and/or segments), and $w_{a,m,j}$ is the estimated loading weight for attribute j, for the ath submodel and the mth repetition. $\overline{w_j}$ (and $\overline{w_{.j}}$) is the average loading weight for all the estimations (for attribute j), and M is the number of repetitions. Note that in Eq. (3) A equals the numbers of bootstrap samples. Please note the similarity between Eqs. (2) and (4), the difference is only the additional correction factor: $1/M$ correcting for the amount of times the Jack-knifing has been run. For the regular leave-one-respondent-out Jack-knifing, the A in Eq. (2) equals the number of respondents (while for Eq. (4) A would equal to 2).

1.3. PLS algorithm – a quick look

In the PLS algorithm, the covariance between **X** and **Y** are maximised through finding a set of scores – $\mathbf{T_X}$ and $\mathbf{T_Y}$, for **X** and **Y** respectively. These are furthermore connected with the variables in each of the matrices through the loadings – $\mathbf{P_X}$ and $\mathbf{P_Y}$. $\mathbf{P_X}$ is often normalised to length 1, while no normalisation is enforced on $\mathbf{P_Y}$. This, of course, has an influence on the size of $\mathbf{T_X}$ and $\mathbf{T_Y}$. It is important to take into account this inherent difference when comparing differences and similarities between A-PLS and PLS-DA, since the scale of the axes cannot be used as a measure of similarity between the two versions of PLS. Visual inspection is better and a more intuitive approach to evaluate the similarity between A-PLS and PLS-DA.

2. Materials and methods

The analyses presented in this paper are based on two datasets from previously published studies with beer as test product (*Dataset 1:* Reinbach et al., 2014; *Dataset 2:* Giacalone et al., 2013). Both studies were conducted on a population of Danish consumers and are briefly described below. Interested readers are referred to the respective publications for full details about products and experimental procedures.

2.1. Dataset 1 (Reinbach et al., 2014)

Seventy-three consumers (46 men and 27 women, aged 18-65) evaluated eight beers chosen to represent distinct sensory profiles. The names of the eight beers are as follows: sea buckthorn beer, fynsk forår, bøgebryg, pine beer, valnød hertug, stjernebryg, enebær stout and thy pilsner. The CATA ballot contained 38 attributes: *BERRIES*, *Blueberry*, *Cranberry*, *Sea-buckthorn*, *Rosehip*, *Other berries*, *FLORAL*, *Elderflower*, *Chamomile*, *Lavender*, *Rose*, *Other floral*, *HOPPY*, *NUTTY*, *Hazelnut*, *Almond*, *Walnut*, *Other nutty*, *ROASTED*, *Roasted bread*, *Caramel coffee*, *Chocolate*, *Other roasted*, *SPICY/HERBAL*, *Juniper berry*, *Bog myrtle*, *Anise*, *Rosemary*, *Cloves*, *Laurel*, *Other*

spicy/herbal, *WOODY*, *Piney*, *Birch*, *Beech*, *Maple*, *Other Woody*. Upper capital names indicate the upper hierarchy of attributes, while the lower capital names are the more detailed attributes. The ballot used in this study required respondents to indicate, for each attribute, whether it applied or not to the evaluated product. This forced-choice variant of CATA is sometimes referred to as "applicability scoring" (Ennis & Ennis, 2013), and is equivalent to classical CATA from a data analytic point of view.

2.2. Dataset 2 (Giacalone et al., 2013)

One-hundred and sixty consumers (92 men and 78 women, aged 18-75) evaluated six commercially available beers, spanning over different beer styles available in the Danish market. The CATA ballot used in this study contained 27 attributes: *Aromatic*, *Beans*, *Berries*, *Bitter*, *Caramel*, *Citrus fruit*, *Dessert spices*, *Dried fruit*, *Flowery*, *Foamy*, *Fruits*, *Fruity*, *Herbs*, *Intense berries*, *Liquor*, *Nuts*, *Refreshing*, *Regional spices*, *Savory spices*, *Smoked*, *Sour*, *Sparkling*, *Spicy*, *Still*, *Sweet*, *Vinous*, *Warming*.

2.3. Data analysis

The CATA data was mean-centred and the information with regards to beer type was converted from a vector indicating the beer number and into a beer dummy matrix with as many columns as there are beers. The value in the column is 1 if the sample corresponds to the specified beer and 0 otherwise. Also this beer dummy matrix was mean-centred prior to analysis by PLS. Dataset 1 was both analysed using an A-PLS approach (predicting the CATA scores based on the beer dummy matrix) as well as the PLS-DA approach (predicting the beer dummy matrix based on the CATA scores). For simplicity of discussion and graphical interpretation all models have been performed using only two PLS components. The results from the A-PLS and PLS-DA were visually compared.

In order to investigate the significance of each of the CATA attributes, three schemes were tested. Two schemes testing different segmentations for Jack-knifing, and one where the uncertainty was estimated by bootstrapping. For Jack-knifing, the common method of performing leave-one-respondent-out segmentation was used, and in addition multiple split-half runs was tested. Each sub-model, both with respect to Jack-knifing and bootstrap sample, was rotated and mirrored towards the model on the calibrated data. This is a common method to ascertain that the estimates are not over estimated (Babamoradi et al., 2013). The uncertainty estimates for loading weights were subsequently estimated based on different amounts of split-half runs, and compared to the uncertainty estimate for leave-one-respondent out.

The importance of the different variables is compared using three different methods: Cochran's Q test, t-test on the Jack-knifed uncertainty of the beta-coefficients (beta-test), and the confidence interval of the loading weights estimated through repeated Jack-knifing (weight-test). The beta-test is based on the smallest p-value across all the beers for both datasets. Cochran's Q test inspects one attribute at a time, while the beta-test and weight-test both are based on models where all the attributes are present. We hypothesise that the beta-test will lie somewhere between the Cochran's Q test and the weight-test. The beta-test is based on a multivariate model, but boils the problem down to a univariate one. For the weight-test the importance is based directly in the uncertainty of the weights in the multivariate space that the loading weights span. This shortcoming of the beta-test also goes for other variable selection techniques such as Selectivity Ratio (SR) (Rajalahti et al., 2009) as well. The main difference with the weight-test and Variable Importance in the Projection (VIP) (Chong & Jun, 2005; Wold, Sjöström, & Eriksson, 2001) is that in VIP the importance of the weights are scaled in accordance to their importance in the model, i.e., a similar difference as between the Euclidean (VIP like) and Mahalanobis distance (weight-test like). In this paper, though, we are not interested in discussing the difference between the variable selection techniques, but rather demonstrate that thinking multivariately leads to a different (and we think better) conclusion than boiling the problem down to a univariate one.

All data analyses were performed in Matlab 7.14 for Windows (Mathworks, Natick, MA, USA, 2010) based on well-known algorithms.

3. Results and discussions

The main part of Section 3 will focus and discuss the results achieved for Dataset 1; only at the end of Section 3.3 will Dataset 2 also be discussed. In order to increase readability we have decided to divide Section 3 in three sub-sections according to their respective focus. To justify the analysis on these two datasets the global test by Meyners et al. (2013) was applied to both datasets. The test showed that there is a significant difference between the products for both datasets.

3.1. Comparing A-PLS and PLS-DA

The first observation which is done upon evaluating the CATA data with A-PLS is that by looking at the **X**-scores the eight different beers just form 8 points in the score space, see Fig. 3. This occurs even though each of these 8 points in the score space consists of a total of 73 different dots (respondents).

This fact, can easily be explained by looking at the algorithm for the PLS (Geladi & Kowalski, 1986), where the first step is to calculate the covariance matrix between **X** and **Y**. Since **X** in this case is a dummy matrix with eight columns which only describes the type of beer, the covariance matrix will only contain 8 different rows (i.e., many rows will be identical, and not the 8 * 73 which it actually consists of). In other words, by performing A-PLS between dummy variables describing a property (e.g., beer type), you get the same answer as from a "collapsed" PCA, where a PCA is run on the average/sum of CATA scores for the different attributes (see Fig. 1B). One problem in Fig. 3, though, is that we have no idea if the difference we are seeing is a significant difference between the products.

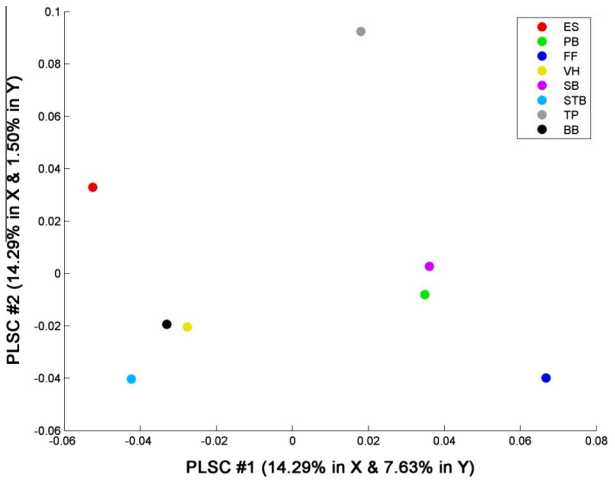

Fig. 3. X-scores from A-PLS coloured according to the eight different beers. Please note that each dot actually is 73 dots (respondents) on top of each other. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

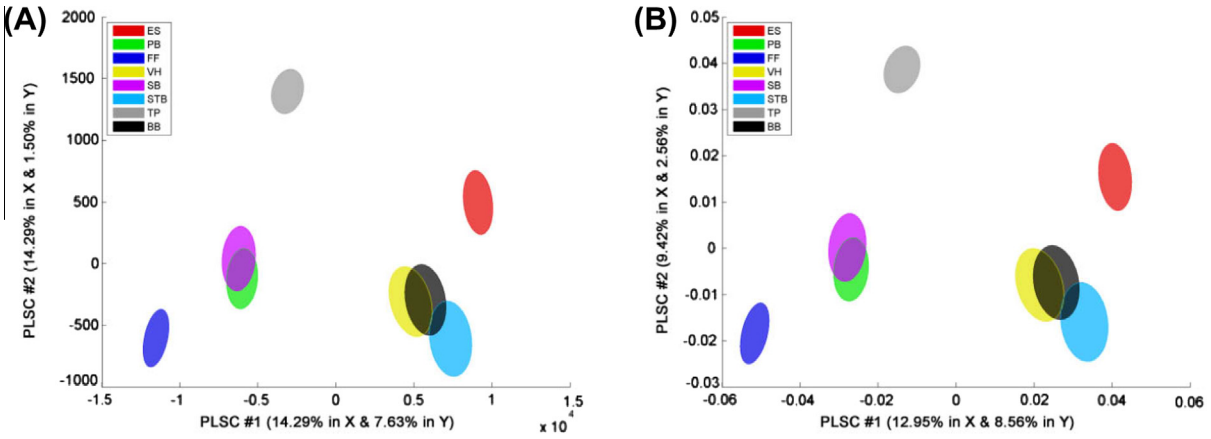

Fig. 4. Comparing how the A-PLS (A) and the PLS-DA (B) estimates the differences between the eight product types. The ellipses show 95% confidence intervals in each direction of variation. Note that these directions do not necessarily lie on the direction of the axes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

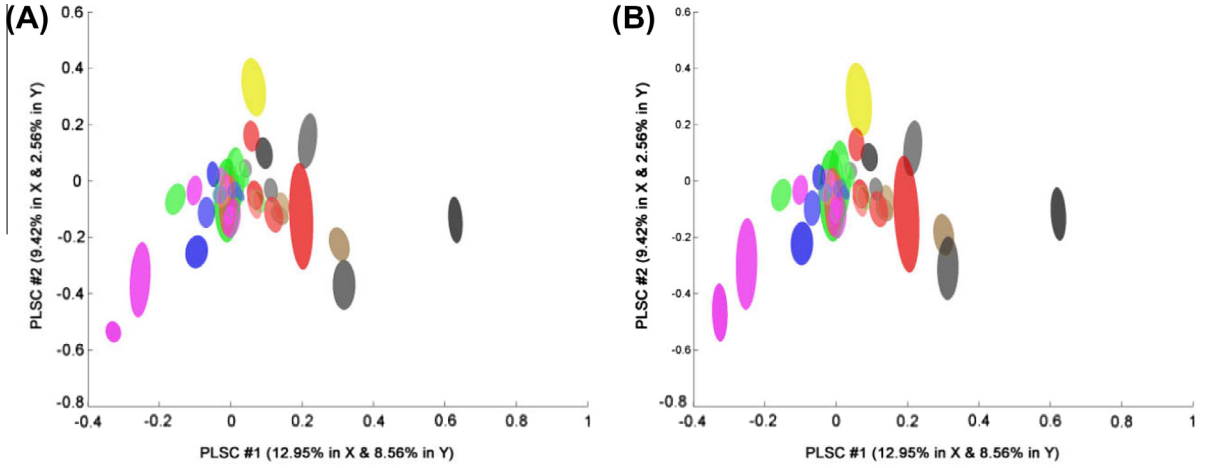

Fig. 5. The estimated confidence limits according to two different ways of calculating them: leave-one-respondent-out (A), and bootstrapping with 2000 runs (B). The ellipses equal one standard error in each direction of variation.

In order to investigate if the respondents have found a difference between the products it is important to investigate the uncertainty in the model. By looking at Fig. 3 it seems that there is no uncertainty between the different respondents and how they perceive the beers (as there are 73 dots exactly on top of each other), but this is, of course, not correct. However, this uncertainty cannot be seen from the **X**-scores ($\mathbf{T_X}$) of the A-PLS as shown in Fig. 3, but one should rather look into the **Y**-scores ($\mathbf{T_Y}$, Fig. 4A). A similar difference between the products is seen if one performs a PLS-DA and investigates the **X**-scores from that analysis (Fig. 4B). At the first glance it may seem like we are showing the same figure twice, but by inspecting the explained variance and the axis it is easily appreciated that these plots are actually from two different models. However, the difference only lies in the scale difference, not in the conclusion with regards to what beers are different from one another; the relative distance between the product types remains the same. In principle, in order to get a good estimate of the uncertainty one should perform some kind of resampling, but as the uncertainty in the product itself is so much larger than any of the respondents, this test can just as well be run on the calibration model itself (the validations do not change this estimate very much). And, as can be appreciated by looking at Fig. 4, the same significant difference between the products can be found both by using A-PLS and PLS-DA.

By inspecting Fig. 4 (you choose yourself which one you prefer to interpret), it is clear that beers like valnød hertug (VH), stjernebryg (STB) and bøgebryg (BB) have similar sensory profiles, while enebær stout (ES) and fynsk forår (FF) clearly are perceived as being very different by the respondents. An even more important observation is that there is no practical difference between the results from A-PLS and PLS-DA. The similarity between A-PLS and PLS-DA was briefly mentioned by Martens et al. (2000) in the same paper where they introduced A-PLS. They state that A-PLS will give information about the sensory attributes, while PLS-DA will indicate what is relevant for the sensory variation. However, as we have shown above, it all depends on whether you are looking on the **X** or the **Y**-scores. We go as far as saying that one of the two is enough. The only "problem", so to speak, is practical: very few, if any, commercial softwares (e.g., Unscrambler) have the capability to show both scores and loadings for both the **X** and the **Y**-block. Most often it is only possible to look at the **X**-block; this is the standard even in dedicated packages in Matlab – PLS toolbox (Eigenvector Research Inc., Manson, WA) and R (R Core Team, 2012) – PLS package (Mevik, Wehrens, & Liland, 2015). This limitation leads to the necessity to perform A-PLS as well as PLS-DA, for the reason given in Section 1. However, as we are basing our analysis on in-house written Matlab code, this problem does not exist. We will therefore select the PLS-DA for the rest of the manuscript.

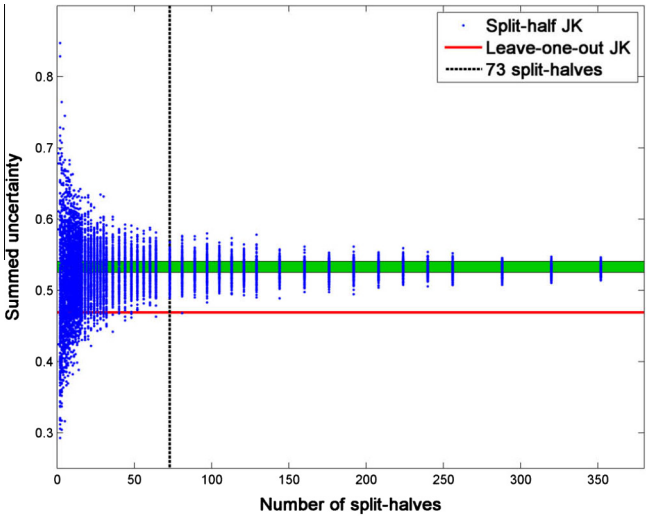

Fig. 6. The evolvement of the summed uncertainty estimate for the 38 attributes. For each number of split-halves 200 random selections from a pool of 800 split-halves have been made. The red is the summed uncertainty for the leave-one-respondent-out case, while the green bar indicates the estimated uncertainty by bootstrap (also tested for 200 different selections of 2000 bootstrap samples). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For most cases you would state that the CATA scores are the independent variables, while the beer types are the dependent ones, and thus a PLS-DA also theoretically makes more sense than A-PLS.

3.2. Uncertainty estimates

Now that we have an idea about which beer types are different, it is of interest to investigate whether there is a difference in how the respondents are using the different CATA attributes to describe the products. As this uncertainty is on the level of the individual CATA-attribute, resampling has to be performed in order to get an estimate of it. As mentioned earlier in the paper, this can be done either through Jack-knifing or a more complex method, such as bootstrapping. The uncertainty estimate from a leave-one-respondent-out Jack-knifing is compared to a 2000 bootstrap sample approach. As is shown in Fig. 5 the leave-one-respondent-out approach estimates slightly too small uncertainties. This effect does not come as a big surprise, as is mentioned by Esbensen and Geladi (2010), but still researchers and users alike wrongly assume that the leave-one-out approach is an appropriate re-sampling method. Fig. 5A and B are quite similar, but if you take a look at the overlap between the different variables, this is clearly smaller in the leave-one-respondent-out, indicating that this approach gives too optimistic uncertainty estimates.

Is it possible to get closer to Fig. 5B by still employing the simpler Jack-knifing methodology? In order to answer this question, 800 split-half Jack-knife runs were performed. Based on these 800 half runs, a number of split-half runs were selected and the total uncertainty estimate was calculated for each selection (Eq. (3) was used in order to calculate this uncertainty). In order to take into account the uncertainty in deciding on the split-half segmentation, the selection of the split-half regimes was done 200 times for each selection of number of split-halves. I.e., for the case with a total of 73 split-half Jack-knifing models there are numerous ways with how this split-half can be performed. From the pool of 800 split-halves, 73 of these have been selected at random and used to estimate the uncertainty. Instead of doing this only once, we have done this a total of 200 times, and thus we can calculate 200 total uncertainty estimates for the given number of split-half runs. The result is shown in Fig. 6.

By inspecting Fig. 6 it can be appreciated that the uncertainty estimate seems to stabilise (i.e., the blue dots are more closely packed) once the number of split-halves approaches the number of respondents. Fig. 5 shows (and as can also be seen in Fig. 6), the leave-one-respondent-out approach clearly under-estimates the uncertainty in the attributes. On the other hand, the multiple split-half Jack-knifing fits nicely with the more complex bootstrap resampling (the green "bar" in Fig. 6).

3.3. Estimating the importance of variables

As mentioned earlier, a common way of estimating the importance of an attribute in CATA analysis is by performing univariate tests, the most common being the Cochran's *Q* test. It was therefore investigated how this test performs compared to the beta-test and the weight-test (both based on a Jack-knife approach). As can be seen from Table 1 there is a general agreement between the three methods. However, as hypothesised, the beta-test lies somewhat in between the two others. Focusing on the attributes where there is a discrepancy between the three methods, we observe that the beta-test agrees with the weight-test 9 times and 5 times with

Table 1
Comparison of significance levels obtained for Dataset 1 by Cochran's *Q* test and two versions of JK-PLS – leave-one-respondent out and 73 split-half runs.

CATA attributes	Cochran's *Q* test	Beta-test[a]	Weight-test[b]
NUTTY	***	***	***
Hazelnut	***	***	***
Almond	***	n.s.	n.s.
Walnuts	***	***	***
Other nutty	n.s.	n.s.	n.s.
ROASTED	***	***	***
Roasted bread	**	***	***
Caramel	***	***	***
Coffee	***	***	***
Chocolate	***	***	***
Other roasted	n.s.	**	**
HOPPY	***	***	***
WOODY	*	n.s.	n.s.
Pine	***	n.s.	n.s.
Birch	***	***	***
Beech	n.s.	n.s.	*
Maple	***	***	***
Other woody	n.s.	n.s.	*
SPICY/HERBAL	***	***	***
Juniper berries	n.s.	**	***
Bog myrtle	**	**	***
Anise	***	***	***
Rosemary	***	n.s.	n.s.
Laurel	n.s.	n.s.	n.s.
Cloves	***	**	***
Other spicy/herbal	n.s.	n.s.	n.s.
FLOWERY	***	***	***
Elderflower	***	***	***
Chamomile	***	***	***
Lavender	**	*	*
Rose	n.s.	***	**
Other floral	n.s.	n.s.	n.s.
BERRIES	**	***	***
Blueberry	*	*	**
Cranberry	n.s.	n.s.	n.s.
Sea buckthorn	*	**	***
Rose hip	**	**	***
Other berries	n.s.	*	*

Significance levels: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; n.s. = non-significant.

[a] The smallest *p* value across the products based on multiple *t*-tests on Jack-knife estimates of the uncertainty in the beta-coefficients from a PLS-DA model. The segmentation used was leave-one-respondent-out.

[b] Based on 73 split-half Jack-knife runs using a PLS-DA model. The significance is estimated based on the uncertainty estimates of the loadings weights.

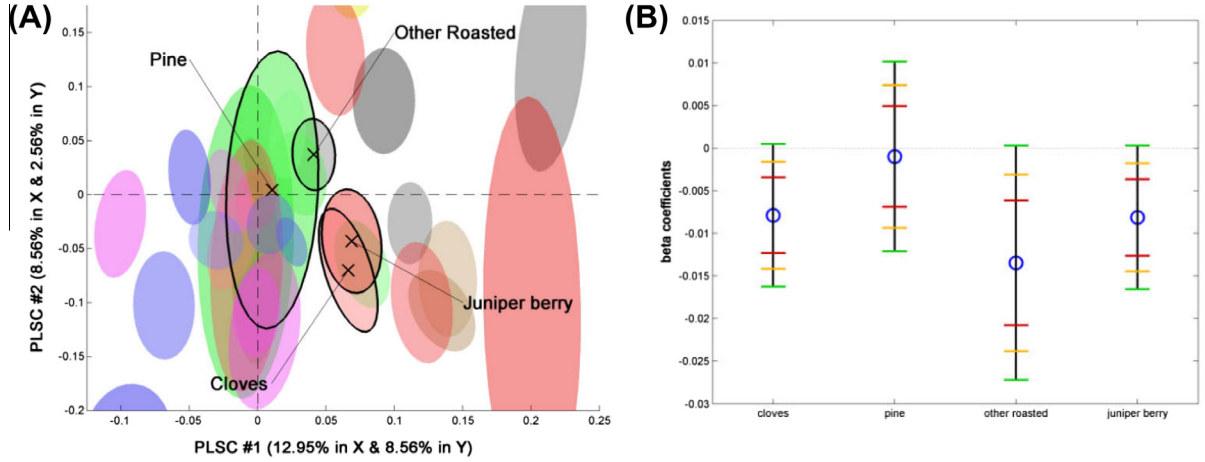

Fig. 7. The uncertainty estimates of the attributes using 73 split-half runs. The focus is on four attributes (cloves, juniper berry, other toasted and pine). (A) The loading weight test, the size of each ellipse is based on one standard error in each direction of variation. (B) The beta-test, the length of the black line equals $p = 0.001$ (as does the green vertical lines). The orange and the red vertical line corresponds to $p = 0.01$ and $p = 0.05$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of significance levels obtained for Dataset 2 by Cochran's *Q* test and two versions of JK-PLSR – leave-one-respondent out and 160 split-half runs.

CATA attributes	Cochran's *Q* test	Beta-test[a]	Weight-test[b]
Floral	***	***	**
Beans	***	***	**
Intense berries	n.s.	n.s.	n.s.
Caramel	***	***	*
Nuts	***	***	*
Savory spices	**	**	n.s.
Dessert spices	*	**	*
Reg. spices	n.s.	n.s.	n.s.
Herbs	n.s.	n.s.	n.s.
Citrus fruit	***	***	*
Berries	n.s.	n.s.	n.s.
Fruit	n.s.	n.s.	n.s.
Dried fruit	***	***	n.s.
Liquor	***	***	*
Bitter	n.s.	n.s.	n.s.
Sparkling	***	***	n.s.
Refreshing	***	***	n.s.
Fruity	***	***	*
Aromatic	***	***	**
Spicy	n.s.	n.s.	n.s.
Still	n.s.	n.s.	n.s.
Smoked	***	***	*
Foamy	n.s.	n.s.	n.s.
Sour	**	***	*
Sweet	***	***	n.s.
Vinous	n.s.	n.s.	n.s.
Warming	***	***	n.s.

Significance levels: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; n.s. = non-significant.

[a] The smallest *p* value across the products based on multiple *t*-tests on Jack-knife estimates of the uncertainty in the beta-coefficients from a PLS-DA model. The segmentation used was leave-one-respondent-out.

[b] Based on 160 split-half Jack-knife runs using a PLS-DA model. The significance is estimated based on the uncertainty estimates of the loadings weights.

Cochran's *Q* test. Only once does the Cochran's *Q* test agree with the loading weight approach. The three occasions the three methods disagree, the beta-test is closer to the weight-test than the Cochran's *Q* test. It seems that the beta-test to a large extent gives a good estimate of the importance of the different variables, also in the multivariate world. But, as the beta-test boils the problem down to a one dimensional *t*-test for each attributes it has its short comings as can be seen from the 9 attributes where they disagree with the weight-test. This can more clearly be seen by inspecting Fig. 7. Here four attributes (cloves, juniper berry, other roasted and pine) have been compared between the weight-test (Fig. 7A) and the beta-test (Fig. 7B). By inspecting these four attributes in Fig. 7, it becomes evident that pine is non-significantly different from zero in the weight-test as well as the beta-test. However, contrary to the three remaining attributes which all are significant at the 0.01 level in the beta-test, two of the three (cloves and juniper berries) are significant down to 0.001 in the weight-test (plot not shown, but the corresponding ellipses do NOT cross the origin). This is mainly due to the fact that the beta-test only looks at the uncertainty at a straight line from the origin and towards the attribute of interest, while the weight-test takes into account how the attribute is distributed in the variable space. The difference is not major (as can be seen from Table 1), but still noticeable and relevant.

For Dataset 2 (Table 2), the difference is larger. Here the beta-test and the Cochran's *Q* test perform very similarly. Only by looking at this table, it would be easy to state that our suggested weight-test gives too large uncertainty estimates and tells us that very little is significant in this dataset. However, we dare say that that actually is the case. By reducing the dataset by selecting only 80 of the 160 respondents the table changes drastically and looks more similar to the results in Table 1, where the beta-test lies somewhat in between Cochran's *Q* test and the weight-test (results not shown). In other words, the big difference in the methods is most probably due to the increased number of samples, and as shown in Section 3.2 the leave-one-respondent-out approach leads to underestimations of the uncertainty. To the knowledge of the authors, there is no current commercial software which allows for such a weight-test, and thus we, again, have to point towards the second best option, namely to use the leave-one-respondent-out strategy.

4. Conclusions

This paper has discussed the possible use of PLS-DA on CATA data, as an easy-to-use and powerful tool for rapid evaluation and interpretation of results. We have shown that A-PLS and PLS-DA gives the same possibilities of exploring the data. However, due to the limitations of most statistical software packages it may still be necessary to perform both. This paper has further explored different validation strategies, and shown that through the use of repeated split-halves the uncertainty in the model becomes more correct, compared to the leave-one-respondent

out, which simply removes too few data points in each sub-model. This comes into play also with regards to the comparison of what method should be used in order to investigate the significance of the different attributes in the study. The most correct result are achieved by estimating the uncertainty in the loading weights by repeated Jack-knife tests. However, due to limitations in current software the best option available for most users is the beta-test based on leave-one-respondent-out Jack-knife estimation of the uncertainty.

Acknowledgements

The datasets used in this paper came from studies funded by the Danish Agency for Science, Technology and Innovation, the Danish Ministry of Economic and Business Affairs, and the Faculty of Science, University of Copenhagen. The work was made possible by research grants funded by University of Copenhagen (author Giacalone).

References

Abdi, H., & Williams, L. J. (2010). Correspondence analysis. In N. Salkind (Ed.), *Encyclopedia of research design*. Thousand Oaks, CA: Sage.

Adams, J., Williams, A., Lancaster, B., & Foley, M. (2007). Advantages and uses of check-all-that-apply response compared to traditional scaling of attributes for salty snacks. *Proceedings of the 7th Pangborn Sensory Science Symposium, Hyatt Regency, Minneapolis, MN, USA*. .

Ares, G., Deliza, R., Barreiro, C., Gimenez, A., & Gámbaro, A. (2010). Comparison of two sensory profiling techniques based on consumer perception. *Food Quality and Preference, 21*, 417–426.

Babamoradi, H., van den Berg, F., & Rinnan, Å. (2013). Bootstrap based confidence limits in principal component analysis – A case study. *Chemometrics and Intelligent Laboratory Systems, 120*, 97–105.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*, 166–173.

Blasius, J., & Greenacre, M. (2014). *Visualization and verbalization of data*. CRC Press.

Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems, 78*, 103–112.

Clausen, S. E. (1998). *Applied correspondence analysis: An introduction*. SAGE Publications.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37*, 256–266.

R Core Team (2012). *R: A language and environment for statistical computing*. 3-900051-07-0. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Ennis, D. M., & Ennis, J. M. (2013). Analysis and thurstonian scaling of applicability scores. *Journal of Sensory Studies, 28*, 188–193.

Esbensen, K., & Geladi, P. (2010). Principles of proper validation: Use and abuse of re-sampling for validation. *Journal of Chemometrics, 14*(3–4), 168–187.

Frøst, M. B., Heymann, H., Bredie, W. L. P., Dijksterhuis, G. B., & Martens, M. (2005). Sensory measurement of dynamic flavour intensity in ice cream with different fat levels and flavourings. *Food Quality and Preference, 16*, 305–314.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta, 185*, 1–17.

Giacalone, D. (2013). *Consumers' perception of novel beers: Sensory, affective and cognitive-contextual aspects* (Doctoral dissertation). Denmark: University of Copenhagen.

Giacalone, D., Bredie, W. L. P., & Frøst, M. B. (2013). "All-in-one test" (AI1): A rapid and easily applicable approach to consumer product testing. *Food Quality and Preference, 27*, 108–119.

Martens, M., Bredie, W. L. P., & Martens, H. (2000). Sensory profiling data studied by partial least squares regression. *Food Quality and Preference, 11*, 147–149.

Martens, H., & Martens, M. (2000). Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference, 11*, 5–16.

Martens, H., & Martens, M. (2001). *Multivariate analysis of quality*. John Wiley & Sons.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometricka, 12*, 153–157.

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2015). URL: <http://cran.r-project.org/web/packages/pls/index.html> Last visited 28.01.15.

Meyners, M., Castura, J. C., & Carr, T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference, 30*, 309–319.

Patil, K. D. (1975). Cochran's Q test: Exact distribution. *Journal of the American Statistical Association, 70*, 186–189.

Rajalahti, T., Arneberg, R., Berven, F. S., Myhr, K.-M., Ulvik, R. J., & Kvalheim, O. M. (2009). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems, 95*, 35–48.

Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L. P., & Frøst, M. B. (2014). Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and Napping®. *Food Quality and Preference, 32*, 160–166.

Rossini, K., Verdun, S., Cariou, V., Qannari, E. M., & Fogliatto, F. S. (2012). PLS discriminant analysis applied to conventional sensory profiling data. *Food Quality and Preference, 23*, 18–24.

Wehrens, R., Putter, H., Lutgarde, M., & Buydens, C. (2000). The bootstrap: A tutorial. *Chemometrics and Intelligent Laboratory Systems, 54*, 35–52.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics, 20*, 397–405.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130.