Principal components analysis of descriptive sensory data; reflections, challenges and suggestions.

Tormod Næs*#, Oliver Tomic+, Isabella Endrizzi** and Paula Varela*

* Nofima, Oslovegen 1, 1433 Ås, Norway

Dept of Food science, Faculty of Sciences, University of Copenhagen, Rolighetsvej 30, 1958 Fredriksberg, Copenhagen, Denmark.

+ Faculty of Science and Technology, Norwegian University of Life Sciences, 1433, Ås, Norway

** Department of Food Quality and Nutrition, Research and Innovation Centre, Fondazione Edmund Mach (FEM),Via E. Mach 1, 38010 S. Michele all'Adige, Italy

Abstract

This paper presents a discussion of principal components analysis of descriptive sensory data. Focus is on standardisation, many correlated variables, validation and the use of descriptive data in preference mapping. Different ways of performing the analysis are presented and discussed with focus on how to obtain informative and reliable results. The results will be commented on in light of experience. All methods will be illustrated by calculations based on real data. The paper ends with a list of suggestions for all the topics covered.

Practical application

The paper is about using PCA in sensory science. The applicability of the methods and ideas presented in this paper are relevant for all types of descriptive sensory data. The ideas are general and comprise areas such as standardisation, validation and many correlated variables. The target group of readers for the paper is the sensory scientist who uses PCA on a daily basis and who may have questions regarding how to use the method the best possible way.

Key words: QDA, PCA, validation, standardisation, partial correlation

1. Introduction

When analyzing data from quantitative descriptive analysis (QDA, see e.g. Stone et al. (2012)), a number of choices are made more or less consciously based on tradition or habits. Some of these choices, however, can have an impact on the solution, and for proper interpretation of results it is important to be aware of their consequences. Special emphasis here will be on the use and interpretation of results from principal components analysis (PCA). Five selected aspects are described briefly below and will be discussed in more detail later in the paper using examples with real data. We emphasise that this is not a exhaustive list covering all possible aspects of PCA.

Aspect 1: Using all individual data or aggregated data

For sensory panels, data contain one intensity score value for each assessor, sample, attribute and replicate. These can be analysed either simultaneously in this initial form, or one can average across assessors and replicates, which is often done in practice. This results in a data matrix with samples as rows and attributes as columns. In this paper we will discuss pros and cons of the two approaches and point at different analysis methods that are suitable in the two cases.

Aspect 2: Standardisation

An important first choice that has to be made when using PCA is whether the variables should be used as they are in their original units or to weight/standardise them in some way. Centring of variables is always done in PCA since interpretation for interval scale data is always easier with a basis at the data centre than in the origin. But how to weigh the relative influence of variables is less obvious.

A common way of making variables comparable is to standardise them to the same variance (obtained by dividing the observations for each variable by its standard deviation), but in many applications this is not done. It is important to stress that standardisation is not primarily a statistical and technical issue, but goes to the core of how to interpret the sensory attributes and to how the assessors are trained and calibrated. In other words, the variability of a sensory attribute is a consequence not only of the difference of the products but also of how the panel is calibrated. If the panel training is properly done, the first two principal components used for visualization - with or without standardisation – will, however, usually coincide quite well if non-significant variables are eliminated. In some cases other types of standardisation than the standard deviation scaling, like for instance Pareto scaling (Eriksson et al. (1999)) may be appropriate.

Aspect 3: Many highly correlated variables

Another choice that has to be made when using PCA is which variables to incorporate into the analysis. Should one use all variables or only a subset reflecting the most important dimensions? If for instance the same phenomenon is described by several variables, the PCA plots may give a biased impression of the relative importance of the underlying sensory dimensions. Obvious examples of this are variables describing the odour and flavour of the same phenomenon and contrasting attributes such as dark/light and soft/hard, but other less obvious examples related to the cognitive or sensing process may also be envisioned. In this paper we will discuss this phenomenon in some detail and give advice regarding what to do in practice. Partial correlation analysis will be proposed as a useful tool in this context. This method may be useful both for making PCA results more relevant to the user and also for obtaining a deeper insight that can lead to improved panel training.

We emphasise that there is nothing wrong with using PCA on the full data set, it will always reflect the internal correlation structure in the whole data set. The potential problem is that the

assessment of the relative importance of underlying sensory dimensions may be biased and sometimes sensory dimensions may appear more/less important than they deserve.

Aspect 4: Validation

Validation is another important issue when using PCA (Næs et al. (2018)). In most applications of PCA one will be interested in knowing to which degree one can rely on the different components extracted. One can of course always consider PCA as only an empirical way of looking at the data, but some assessment of confidence in the components is also often wanted. In this paper we discuss a number of ways of how this can be done. Different types of validity will also be discussed.

Aspect 5: QDA used in relation to consumer data

In some cases, not all sensory attributes are important for the purpose they are used for. An example is preference mapping, where for instance a certain spice or salt level may be important for consumer preference, but its effect is blurred by the presence of a large number of attributes that are irrelevant for this problem. If for instance only two principal components are considered in external preference mapping, the effect of a single important variable appearing in the third component may pass unnoticed. Another example is studies of satiety, where in most cases only the texture attributes will be relevant (Nguyen et al. (2019)), not the whole sensory profile.

The present paper is a discussion of these five aspects with focus on interpretation and what type of effects they may have on the results. Both personal experience, concrete results from sensory data and basic principles will be important in the discussion. The main purpose is to provide guidelines for the sensory analyst in industry and science and suggestions of how to use PCA in a safe and reliable way. The paper is not intended for the specialist statistician, but for the more typical users of these methods in their daily activities and practice. Some

possible pitfalls are underlined and some new suggestions and tools will be presented and discussed. A short introduction to PCA is provided here, but for a a thorough description of several more aspects of PCA we refer to Jolliffe (2010). At the end of the paper (Section 10) a number of conclusions and recommendations are given for each of the issues discussed. The phenomena discussed will be illustrated by examples using real sensory data sets.

2. Structure of descriptive sensory data

The focus of the present paper is the use of PCA for descriptive sensory data (QDA data). In most cases the entries in such data sets will lie between a lower and an upper limit on some sort of intensity scale. The different attributes are calibrated to be positioned within this interval. It should be mentioned that although PCA is a very important tool in this context, a proper analysis and interpretation of each of the attributes separately is always recommended.

For the purpose of interpretation and also for some of the tools proposed, the sensory data will be thought of as generated according to an experimental design with assessors and products as the two factors in the design. In more technical terms, each sensory variable can be considered a sum of contributions from the two factors, product and assessor, i.e.

$$y_{ijr} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijr} \qquad (1)$$

where y_{ijr} is the measurement for product i (i=1,…,I), assessor j (j=1,…J) and replicate r (r=1,…,R). The α represents the product effect, β the assessor effect, $\alpha\beta$ the interaction between the two and ε represents the random error. Note that when the samples are obtained according to an experimental design, one can replace the samples effect α by separate effects for the design factors (see e.g. Næs et al. (2018)). It should be mentioned that for ANOVA purposes, more sophisticated models than (1) have also been proposed (Brockhoff et al. (2015)).

If we combine the models in (1) for the all sensory attributes (K), the joint model can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \qquad (2)$$

where **Y** is the matrix of sensory data (each column of **Y** represents an attribute), the **X** is a dummy matrix (containing zeros and ones) representing the design, **B** is the matrix of unknown regression coefficients and **E** is the random error, i.e. the variation in **Y** not accounted for by the design. The different columns of **B** represent the coefficients for the different sensory variables, i.e. they correspond to the Greek letters in Equation (1). The number of columns/attributes in the data matrix **Y** is K and the number of rows will be equal to I*J*R (products*assessors*replicates). We refer to Figure 1a for an illustration of the data structure in Equation (2). Some places below, the data set **Y** without any prior modifications or transforms will be called the raw data.

The data can be analysed by PCA directly using **Y** in Equation (2) or using the data matrix obtained after averaging across assessors and replicates. In this case **Y** is sometimes referred to as a consensus matrix and consists of I rows and K columns.

Another way of organising QDA data is by using a three-way array structure with the rows corresponding to samples*replicates, columns to attributes and slices to the different assessors (Figure 1b). This type of data structure can be analysed by so-called multi-way methods such as PARAFAC (Bro et al. (2008)), or one of the Tucker methods (Tucker (1964)), which are extensions of standard PCA. The data set organised as in Equation (2) is referred to as a three-way data set which has been unfolded (See Figure 1b) vertically. The data structure to the right in Figure 1b corresponds to **Y** in Figure 1a and Equation (2). The three-way structure and analysis will not be pursued further here.

3. Short description of PCA.

Principal component analysis is a so-called component method. This means that it is based on the idea that a large number of variables in **Y** can be approximated by a small number of so-called components **T** (sometimes called axes or latent variables) calculated as linear combination **YW,** where **W** is the matrix of so-called loading weights (columns of **W** have length= 1). The components are found by maximising their variance and such that each new component extracted is orthogonal/uncorrelated with previous ones. The first component describes the most of the variability, the second is the next in the order etc. A consequence of the criterion used is that variables or variable groups with large variance will have a stronger impact on the solution than the rest. Usually one extracts only a few components treating the rest of the variability as noise. After calculation of the components, they can be related to **Y** by regression in order to find the loadings **P**. The model for PCA can be written as

$$\mathbf{Y}=\mathbf{T}\mathbf{P}^{\mathbf{T}}+\mathbf{E} \qquad \mathbf{(3)}$$

Here **T** represents the few components extracted to approximate **Y** and the **E** is usually thought of as noise. The **T**'s are called scores and the **P**'s loadings and are usually plotted in scatter plots for interpretation of results.

Although there is an arbitrary choice related to the scaling of **T** relative to **P**, one usually organises the solution such that the length of the loading vectors, columns in **P**, is equal to 1. Then the variance of the columns of **T** represent variability along the unit axes defined by the loadings. The components and loadings can be found using the singular value decomposition (SVD), which is a standard mathematical tool for decomposing a general matrix. For a thorough introduction to PCA we refer to Jolliffe (2010). In this paper we will consider the components in the order they appear according to explained variance and no focus will be on rotations.

4. PCA for original or averaged data?

Averaged data for studying product differences.

In most cases in the literature, panel averages are used both for interpretation and for estimating relations with other data, for instance chemical data. This is a sensible strategy if focus is on product differences, but should always be accompanied with proper checking of the panellist quality. If an assessor is clearly outlying/different, it is questionable to keep him/her as a part of the analysis. This is in particular true if the number of assessors is low since in such cases outliers may have a larger impact on the analysis. A number of methods have been developed for the purpose of checking panel performance (see e.g. PanelCheck software, Dijksterhuis (1995), Tomic et al. (2007), Tomic et al. (2010), Dahl and Næs (2004, 2009)) and Dahl et al. (2008), Tomic et al. (2013)).

Different types of panel averages

It should be mentioned that there are different ways of obtaining panel averages (or a panel consensus). One of them is to use straightforward averaging as will be focused here. Other possibilities are Generalised Procrustes analysis (Gower (1975)), STATIS (see e.g. Schlich (1996)), multiple factors analysis (MFA, Escofier and Pages (1995)) and various scaling techniques (Romano et al (2008)). Generalised Procrustes analysis rotates, reflects and scales (isotropic scaling) the individual assessor data matrices to make them as similar as possible and then afterwards calculates the consensus as the average. The STATIS method calculates a weighted average of the individual (cross-product) matrices, where the weights depend on the RV coefficients between them. MFA concatenates the individual data matrices horizontally and essentially runs a PCA on the combined matrix after a specific individual scaling of each of them. The resulting scores matrix of this PCA is then used as a consensus for the individual assessors. An alternative to MFA, with a similar underlying idea is the Tucker-2 method used in Dahl and Næs (2009). The scaling methods in Romano et al. (2008) are used to eliminate additive and multiplicative differences among assessors before averaging. Note that all these

methods are also suitable for investigating individual differences among assessors (See e.g. Næs et al (2018)).

PCA for original data

If focus is also on individual differences between assessors, one can use the original **Y** data in (2) directly without averaging. There will be several more points in the score plot, one score for each replicate, assessor and sample combination. For improved interpretation one can include colours and sample averages as will be illustrated here. This plot can be useful for visualising differences/disagreement among assessors.

If the assessor points for each sample deviate strongly from each other, it provides evidence that the assessors disagree to a larger extent. But in general, the differences will always look quite large in this case due to noise and different use of the scale. For this reason, it is also possible, to centre (and also standardise) each of the assessor data matrices before PCA. By doing this one eliminates differences in intensity level on the scale between assessors before analysis (see also Romano et al. (2008)).

Note that the explained variances when using the original data will normally be smaller for the original data than for the averages since averaging reduces noise (see also example below).

If focus is only on product differences, we recommend to use averaged data because of simpler plots.

5. Standardisation

Different practices for standardisation in PCA exist, but whether to do it or not may sometimes seem to be more a matter of habit than of serious reflection and consideration. The issue of standardisation is important both for panel averages and for individual data.

For PCA in general, many different types of standardisation are used, but here we confine ourselves to the most used namely division by standard deviation. It should be mentioned that using PCA on standardised data is what some authors phrase as using the correlation matrix as the basis for the calculation of components.

Standardisation is not primarily a statistical issue

It is important to emphasize that standardisation is not primarily a statistical issue. Whether to do it or not is strongly related to how the sensory attributes are calibrated and interpreted. This is clearly a decision with a subjective element, made by the panel leader or agreed upon by the panel during the training session. One could easily envision that two panels with the same sensitivity to product differences could be calibrated in a different way leading to another ratio between the variability of for instance sweetness and hardness and then possibly different PCA results. Culture and context will also have an influence on this matter, which can lead to different plots and varying interpretation of results.

The complexity of the attributes will play a role (i.e. training and calibration on complex attributes as for example creaminess is not straightforward), as well as the variability of references. Taste and flavor attributes are usually easier to anchor with reference solutions or products as compared to texture attributes.

A crucial question is whether one can justify that two attributes, possibly representing different modalities, can be compared directly or not. Let us for instance consider two non-standardised variables hardness and sweetness, the former with standard deviation equal to 1 and the other with standard deviation equal to 3. From this it seems that the variability of hardness is 3 times larger than the variability of sweetness. The question is how to interpret this in an appropriate manner. Can variability in hardness and in sweetness really be compared this simply?

Interpretation of PCA with and without standardisation

If no standardisation is done, the rationale is that the ratio of the standard deviations of the attributes is considered meaningful. In other words, without standardisation, one relies on the meaningfulness of the subjective decisions made in the calibration phase. A consequence of this is that the variables with the larger variance will have the strongest influence on the PCA solution.

If on the other hand the variables are standardised by their standard deviation (or span or other multiplicative constants), the relative differences in standard deviation are disregarded. This corresponds conceptually to saying that for each of the attributes, the anchors (defining the span) used for calibration of the different attributes are placed approximately at the same place on the scale. This implies that differences between two samples are always interpreted relative to the same variability or span. This means that variables with for instance initial standard deviations equal to 1 and 3, will end up being compared as though they have the same standard deviation.

It is important to mention that when using standardisation, the variance of all variables will be the same. This implies that only the number of variables related to a sensory dimension will be the driver for order of the components. If for instance one phenomenon is described using four highly corelated sensory attributes and another phenomenon is represented by one attribute only, the first principal component will represent the phenomenon with the four attributes and the second component will represent the other variable. Therefore, in such cases, importance of dimensions (in terms of explained variance) is driven by the number of correlated attributes representing the same phenomenon rather than by the most dominating

sensory dimension. This shows that it is not obvious how to define the concept of common concept of ‘most important sensory dimensions’ using QDA and PCA

Eliminate non-significant attributes

If one decides to standardise the data, it is important to recognise that variables with very small variability will then be comparable (i.e. have the same influence) to the rest. A possible problem with this is that variables containing mainly noise may become important in the analysis and results. A pragmatic approach to avoid this problem is to test all attributes for significant product effect, using ANOVA based on the model (1) above, or a more sophisticated model as proposed in Brockhoff et al. (2015). If an attribute is non-significant, the variable should be disregarded, thus reducing the amount of noise in the data. It is important to emphasise that this approach should be used with care since significance of a variable is not an objective concept and that significance of an attribute can be deflated due to a few of the assessors only. Another aspect of eliminating non-significance variables is that variables with low significance are eliminated and one is left only with variables which have already proved their significance in the data. Generally, it is our view that, it is most often better, from a pragmatic point of view, to remove non-significant variables in order to avoid further problems with noisy attributes.

Using correlation loadings plot

Correlations loadings (Martens and Martens (2001)) are defined as the correlations between the original variables and the components. This provides a plot similar to the standard loadings plot with two axes, but is in addition most often equipped with circles indicating 100% and 50% explained variance. The correlations loadings have the advantage that they highlight variables with low variance that may have a strong correlation with the components.

It is tempting to think of correlation loadings as a way of eliminating the problem of standardisation. However, this is not always the case since correlation loadings only represent a post processing procedure after the principal components have been estimated. The method may be better at highlighting the relations between variables with a small initial variance (and which therefore have little influence on the solution) and the components, but this does not change the data for which PCA is calculated. For standardized data, the two are the same except for a scaling factor. We here use the unit circle scaling for the correlation loadings.

6. Correlations between variables

A PCA solution is determined by the variance-covariance structure among all the variables in **Y**. More precisely, PCA tries to explain as much as possible of the variance in **Y**. This means for instance that if several variables describe the same phenomenon, this phenomenon may represent more variability than the underlying phenomenon deserves, possibly only because a panel leader may have chosen to have the panel evaluate these variables. To PCA it will then look more important than other dimensions which may be represented only by one single attribute.

Avoiding highly correlated variables

It is generally recommended that too much repetition of information should be avoided in order to reduce unnecessary bias and focus for the PCA. Some of these repetitions may be quite obvious such as using confounding attributes as for example dark/light and hard/tender (see introduction), while others may be more subtle and difficult to identify directly without data analysis. Assessors may for instance have problems discriminating between two or more cognitively similar attributes and will automatically score them similarly. This is known as halo dumping effect. It comes from the human desire of consistent cognitive structures and has been widely described in the sensory literature (see for example Clark and Lawless

(1994)). Correlation between unrelated attributes may also happen when one salient negative attribute causes another to be rated in the same direction, Such correlations are known as horn effects, common when describing defective samples (Lawless and Heyman (2010)). This is an unfortunate situation and having tools to detect such cognitive coincidence is important for more relevant analysis and interpretation of PCA and for improved training of the panel. One of the objectives of panel training is to achieve de-correlation of the attributes, and avoid redundancy leading to particular issues in multi-product panels, as some attributes can be correlated for one product but not for another.

Correlations at different levels

Correlation between attributes/columns in **Y** can be due to correlation induced by the design (**X** in Equation (2), representing sample, assessor and interaction) and by the random error **E** in the model. The correlations between variables in **XB** are the most important since these are functions of the design of the study. Correlations among the variables in **E** are, however, conceptually more problematic. This calls for investigating the correlation structure for **XB** and **E** separately and sometimes also for the products and assessors separately. We will next discuss a possible tool to use for detecting correlations among the variables in the before we describe briefly a few methods for studying **XB** by PCA.

Partial correlation for detecting correlations among random errors in equation (2)

The concept of partial correlation between variables was developed for the purpose of correlating two variables with each other after they have been conditioned upon a third variable (or set of variables). This is equivalent to correlating the residuals **E** for the two variables with each other after they have been regressed onto the same variables. If the partial correlation among two variables is high, one should consider eliminating one of them from the PCA to avoid the problem discussed above. This type of information may also be

important for retraining the panel and to improve its performance. Since this type of correlation will most typically be present at the individual level, correlation between residuals at an individual level will be given the strongest focus here.

There are different ways of implementing this idea, but here we will confine ourselves to results obtained from the residuals for all variables after a full two-way ANOVA of the data (Equation (1)). The true partial correlations will be presented, but for the individual assessors we will only consider correlations between the residuals from the full ANOVA of all assessors.

PCA for the systematic part **XB** *of equation (2)*

An important PCA based methods for analysing the systematic part **XB** is ASCA (Jansen et al (2005). PCA plots for this method can be used to reveal cases with highly overlapping attributes as discussed above. The effects of the assessor and product (and their interactions) are first estimated using the model (1) and standard ANOVA methods. Then the effects for the different factors are further analysed by PCA using all the response variables. This is equivalent to estimating **B** in Equation (2), then splitting the **XB** contribution into three parts, the assessor part, product part and the interaction part. Analysing each of them by PCA results in three separate PCA models. In mathematical terms this means that **XB** is essentially written as $\mathbf{X_1B_1+X_2B_2+X_3B_3}$ and each of the terms is treated separately by PCA after estimation of the **B**'s. In this way information is obtained about the variability structure of the sensory attributes for the assessors, products and interactions separately (see Liland et al (2018)). This means that this method can reveal correlation structure at the sample level and assessor level separately. The PC-ANOVA (Luciano and Næs (2009)) is related, but reverses the order of ANOVA and PCA. First a PCA is run for **Y** and then the scores for the first few components are related separately to the design using the model (1).

7. Validation of PCA models

When using PCA, there is always a question of how many dimensions/components that can be interpreted safely, regardless of whether it is applied to individual assessor data or panel averages. PCA will always provide a model or solution, but the question is whether it is valid in the sense that it is reproducible. Before considering methods for assessing validity, we will discuss different types of validity.

7.1. External validity.

This validity looks into whether the model can tell something about a larger population of samples or not. In sensory science this case is often not of highest interest since the samples considered are the samples at hand and very often these are not selected to represent a larger population. Typically, the samples are from product development, quality control or another more specific situation and as such, the samples do not represent something else than themselves and the perceptual space they span. The fact that the number of samples is often also very small and sometimes based on an experimental design, makes it even more difficult to interpret them as representing something bigger.

Leave one-out cross-validation (CV) of samples is a method which was originally developed for external validation of regression models (Stone (1974)). It can also in principle be applied for PCA if the explained variance of **Y** is used as a criterion. As argued among others in Næs et al. (2018), this method is for the above reasons not always suitable in PCA studies of sensory data. It may give reasonable indications of number of components to rely on in medium size data sets, but one should, always be careful with small data sets (for instance 4-5 samples) , especially if the samples were designed to be very different from each other . In the results section we will give an example for a very small data set and a normally sized set.

For standardised data, the leave-one-out CV can be done in slightly different ways. Here we have used the following procedure: every time an object is left out, the remaining data are standardised prior to PCA. Then the sample which is left out is corrected for the mean and the standard deviations from the samples used for model building, before calculating how well it fits.

7.2 Internal validity.

Internal validity of a component means that a component is more meaningful or describes a larger percentage of variance than the variance that can be obtained by chance, i.e. in data sets without an underlying structure. Therefore, comparing true explained variance with what is obtained by chance is a possibility. This type of validity is only referring to the data set under study and will not tell anything about how well the model represents a population of other samples. The cross-validation as defined by Wold (1978), which is based on successively creating subsets for validation by eliminating entries according to a diagonal pattern of the data set, can be considered an internal validation method. Here we will, however, concentrate on a method based on permutations as proposed in Endrizzi et al. (2014) and later studied and modified by Vitale et al. (2017). We will here use the original version.

Permutation testing

The idea behind the method is that for each new component to be tested, the residuals from the model based on all previous components are permuted (for each column separately) and then orthogonalised with respect to both columns and rows (since this is the case for the true residuals in a PCA). Then, one calculates the explained variance of the permuted residuals data set and compares it with the true explained variance. This is done by comparing the explained variances for the component considered relative to the variance left in their

411 respective data sets (permuted residuals and true residuals). The procedure is repeated for a
412 large number of permutations (for instance 1000, as used here). The results are then presented
413 in a plot with component number on the X-axis and the explained variances as described
414 above on the Y-axis. For the real data, there is only one point for each component, but for the
415 permuted data, we will here present three values, the median, the lower 5% percentile and the
416 upper 5% percentile, obtained from a large number of permutations. The lower and upper
417 values are there for assessing the uncertainty of the estimates. If the true value falls clearly
418 above the confidence band obtained by the two percentiles, the component can be judged
419 significantly different from that generated by chance and therefore worth looking at. Although
420 assessing the number of components is essentially a one-sided test, we here prefer the setup
421 used to indicate the uncertainty in both directions. For details we refer to Endrizzi et al.
422 (2014).

423 *Assessor based cross-validation*

424 If original data are available at individual assessor level, another possible internal validation
425 method is to compare results for the different assessors, i.e. to cross-validate the assessors
426 instead of the samples. We here refer to the block splitting according to assessor illustrated to
427 the right in Figure 1b. A possible way of doing this is to project each assessor, i.e. each
428 segment removed, onto the space spanned by the rest of the assessors and compute the
429 average explained variance over the segments. This method can also be used to identify
430 outlying assessors by looking at the individual contributions to the explained variance.

431 **7.3 Validation using external information.**

432 In some cases, there may be other data available about the samples, for instance chemistry
433 data, spectroscopy data or simply the experimental design. In such cases it is possible to
434 regress the (for instance) average sensory attribute scores (across assessor and replicates) onto

the external data and then evaluate how much of the sensory data that can be accounted for by the external variables/measurements. Such a method was used in Dahl and Næs (2004) for relating the average sensory profile to external near infrared (NIR) spectra. Explained variance of the sensory profile obtained from the NIR data was then used as criterion of validity. In the paper the same was also done for each individual assessor separately in order to identify outliers.

If PCA is run on the raw data **Y** (equation 2), the PC-ANOVA method mentioned above can also be used for validation. Each principal component for the full data set is now regressed onto the design variables (product, assessor and interactions) using the model (1). Note that this can be done in all possible cases with more than one replicate since the sample factor here only refers to the samples tested and not necessarily to a particular experimental design for the samples. It must be stressed, however, that the significance tests in such a model may be quite strong tests due to the large number of observations. One should therefore in addition to looking at degree of significance also look at the explained variances of the components in order to evaluate relevance. A component with very small explained variance and only borderline significant product factors is usually not worth focusing on too much. Significance testing in this case may therefore in general be more useful for assessing the significance of the first 2-3 components rather than evaluating how many components further out that are significant.

7.4 Validation using confidence intervals.

In addition to focusing directly on the significance of a component, confidence intervals or ellipsoids for each sample is a good option. They are primarily meant for assessing stability of solutions, but can also be useful for indicating how many components that are worth considering. Bootstrap procedures as illustrated for instance in Cadoret and Husson (2013) are the most important to use in this case. The method is based on resampling assessors at random

(the same number as in the original panel) and calculating the scores for each selection (after averaging over assessors). These are then projected onto the scores plot of the original averaged PCA and confidence ellipses are drawn based on this for each sample.

8. Implications for relations to consumer data

As mentioned in the introduction, very often a sensory data set is not only used for understanding the variability in the sensory properties of samples. A typical example is preference mapping where the main focus is on relating consumer liking to sensory data. One can do this by analysing one sensory attribute at a time, but a more typical way is to use PCA of the sensory data (or PLS regression) and regress the liking for different consumers onto the first couple of components (often only 2). If then a specific attribute with minor relation to the main variability of the sensory data set, has an important influence on the liking, it will not be visible in standard external preference mapping analysis with 2 components. Typical examples are salt level and spices which may influence liking strongly, but don't account for much variability in the sensory data. One should therefore inspect more than 2 components or supplement (or replace) the analysis with an internal preference mapping, where PCA is applied to the liking data and sensory data are regressed onto the these principal components. PLS regression could be another alternative for such data (see e.g. Næs et al. (2018).)

Satiety studies is another important example where the whole sensory profile is not needed for explaining consumer data. This was demonstrated in Nguyen et al (2019). In such cases, the texture properties are the essential ones for relating to satiety; the rest may not add information to explain the problem at hand, or can at worst blur the focus and results of the study.

9. Case studies

9.1 Data sets used.

Table 1 shows the structure of the 3 data sets used in the different examples.

9.2. Case 1. Should one average or not before computing PCA on sensory data? Exemplified using yogurt data.

The data used for visualizing the differences between using the PCA for average data and for the individual data before averaging is a yoghurt dataset with 8 samples and 21 attributes, (Nguyen et al. (2019)). An experimental design with 3 factors at two levels is used for producing the samples. In this case we focus on standardised data for visualization (after elimination of the single non-significant attribute at 5% level).

The results are presented for panel averages and raw data in Figure 2 and Figure 3.In Figure 3, the average component scores across assessors for each sample are superimposed using diamond shapes. As can be seen, the loadings are quite similar for the two PCA models, but the explained variances are larger for the averaged data due to the averaging process, as explained above. The main difference in loadings is that dryness in mouth and astringent form an own group of attributes for the individual data while for standardised data they are grouped together with sandy, stale odour, etc. There are quite large individual differences around each sample average in Figure 3 (scores with same colour). Still, the average scores for each sample are quite similar to the scores in Figure 2. This means that the essential information is similar for the two analyses. The former provides a simpler plot, while the second gives an opportunity for studying individual differences. As will be seen below, the latter also allows for an ANOVA test for the components. In practice choosing between the two is often a matter of scope of the study and need for simplicity. Most of the discussion below will be focused on average data.

9.3 Case 2. Should one standardize or not before PCA? Exemplified using olive oil data.

An illustration of the effect of standardisation will be given using data from sensory analysis of olive oil (based on averages over assessors). The results are presented in Figures 4a, b, c and d. Figure 4a gives results from PCA on the full set of variables without standardisation, while in Figure 4b, PCA is based on the full set of standardised variables, Figure 4c shows results of PCA for only significant variables, not standardised, while Figure 4d shows PCA results for significant standardised variables. In all cases the explained variances were high, about 90% after 3 components. The three components look significant using leave-one-out cross-validation, and this is also confirmed by the other premutation based method to be shown below.

The Figure 4a shows that loadings and correlation loadings plot are quite different without standardisation. The Figure 4b shows that the scores plot change significantly after standardisation, but now the loadings and correlation loadings are quite similar. Correlation loadings are also different in Figure 4a and Figure 4b. This means that standardisation has an effect on scores and loadings if used on all variables without considering significance. Also, correlation loadings may change with standardisation.

After eliminating non-significant variables (Figure 4c. 6 attributes eliminated), we see that the scores are back again to the ones obtained without standardisation for the full set of variables (Figure 4a). Correlation loadings and loadings are still different, but less so if we compare with the full data set. Standardisation (Figure 4d) now has little effect (for reduced data) on the loadings except for one variable close to the middle. Scores are almost the same for Figure 4c and Figure 4d. After standardisation, loadings and correlation loadings in Figure 4d are identical except for the scaling.

In conclusion. After elimination of non-significant variables, the results are similar regardless of whether one standardised or not. This is true for both scores and loadings.

Comparing full and reduced data sets, we see that scores are almost the same except for the standardised full data set (Figure 4b). Two of the attributes (acidic-O and oxidised-O) that show up in the full data set along the second component are not present in Figure 4c and Figure 4d since they are non-significant. They are also less visible in Figure 4a. These two are examples of variables that are 'inflated' when standardised. This phenomenon is quite frequent with off-flavours or other attributes that may appear in low intensities (i.e. spicy). After standardisation low scoring attributes will get a larger importance in the outcome.

Our advice is to eliminate non-significant variables since it then matters less what is done regarding standardisation. The standardised results with all variables, including non-significant ones, are the most different from the rest. One should focus on a good training for the low scoring attributes when relevant for the products or objective of the study.

9.4 Case 3. Many correlated sensory variables. Exemplified using yogurt and olive oil data.

Figure 2 shows PCA results from the yogurt experiment in Nguyen et al. (2019) (based on a 2^3 design). Most of the variables contrast each other along the first axis. This means that the large variability accounted for along this axis to a large extent is due to the many variables measuring more or less the same phenomenon. This is important information per se, but it clearly gives a biased impression of the relative importance of the two components or underlying dimensions (62% and 20%). Eliminating several of the highly correlated variables along the first component, leads to a different relative weighting of the two axes. In other words, the relative importance of the components is dependent on how many strongly correlated variables that are in the data set.

In practice there is no fixed rule for how to possibly reduce the profile other than the obvious ones, for instance dark/light. It is, however, important to be aware of this fact and interpret results accordingly.

Partial correlation results

An illustration of the use of the partial correlation concept discussed above is given in Figure 5 for the olive oil data set, both for the whole panel (Figure 5a) and for three individual assessors (presented in Figures 5b, 5c and 5d). There is some correspondence between panel and individuals, but the individuals are also quite different. The panel clearly has a large partial correlation between grass flavour and grass odour, between astringency and burning, between astringency and bitter and between bitter and burning. The same tendency holds for two of the individuals presented, but the third does not share this particular tendency. For the assessor in Figure 5b, there are also many partial correlations among some of the attributes in the middle of the plot, for instance between grass flavour and a number of the other attributes. For this specific assessor there is good reason to question his/her interpretation of the attributes involved and consider a retraining.

9.5 Case 4. Validation based on cross-validation and permutation testing. Exemplified using olive oil data

Figure 6 shows results from the permutation test (a) and standard leave-one-out cross-validation (b) for the olive oil data (see above for details) In the permutation test the true explained variance is far outside the confidence interval for components up to 3. After that it is inside, which indicates that from component 4 one cannot distinguish the component from noise. Ten components is the maximum number possible and therefore no confidence interval can be computed for the tenth component.

This data set is also quite suitable for the leave-one-out CV since there are many very similar samples and no unique ones. As can be seen (based on the explained variance along the vertical axis), also the CV indicates clearly that at least 3 components can be interpreted. After that the improvement is negligible. The advantage of the randomisation test is that it gives a statement of significance.

An illustration based on reduced data

For illustrating the problems with standard leave one out cross-validation for small data sets, we selected a subset consisting of only 4 samples from the olive oil data and computed a new PCA model based on standardized data. The scores and correlation loadings are given in Figure 7a) and Figure 7b) respectively. The leave one out CV (Figure 7c) gives meaningless results since each sample is unique and the model changes substantially every time one sample out of four is left out during cross-validation. Note that a negative value of explained variance is not possible when fitting the data by PCA, but for validation it can happen when data left out (a segment or single samples) fit very poorly to the model estimated by the rest of the data.

The permutation method (Figure 7d), on the other hand, indicates that the first component is reliable, while the second is not. This means that the vertical axis has no statistical power regarding interpretation. In other words, there is no general tendency (underlying common component) representing common variability among samples along the second component. It should be emphasized, however, that statistical properties of the permutation test for such small data sets have not yet been tested out, so care must be taken not to overinterpret the results. It should also be mentioned that this is a very extreme case for CV and incorporated just to illustrate how problematic it can be for very small data sets.

An interesting observation is that the loadings plot change when a subset (oils 3, 7, 10 and 11) of the full set of samples (oil 1-11) is used (see Figure 4d). This underlines that interpretation of a subset of samples only relates to this specific subset at hand and cannot be generalised to the sensory space of the full set of samples. Conclusions will then always be local and of limited value for saying something about a larger set of 'similar' samples.

The use of PC-ANOVA for validation

PC-ANOVA (Luciano, G. and Næs, T.(2009) was applied to the standardised yogurt data and compared to the use of the permutation test for the consensus/average data set. The results are presented in Figure 8 and Figure 9. As can be seen, the results correspond reasonably well, the first three components are obviously significant, while number 4 is more questionable. It seems that the PC-ANOVA finds significance further out (components 5 and 6), but these components represent so small variance that they are not very interesting in practice. Also, the fact that component number 4 is non-significant is an indication that one should not consider further components after component 3. The explained variances for the 5 first consensus components are 64.4, 21.1, 9.5, 2.7 and 1.2. For the PCA done on raw data the corresponding values are 28.2, 17.2, 10.4, 9.1 and 6.8. As can be seen, the drop in this case is smaller from the fist to the second component.

9.6. Case 5. Relations between QDA and consumer data. Exemplified using bread data.

For this example based on external preference mapping, a bread data set with 8 samples (based on a 2^3 design) and 13 attributes is used. The data set consists of both QDA data and consumer liking of the same samples. Only the averages will be considered for QDA.

In Figure 10 correlation loadings plots of component 1 vs. component 2 and for component 1 vs. component 3 are shown. As can be seen, there is a major tendency in liking towards

component 3 dominated by salt taste. This tendency is not visible in the plot of component 1 vs. component 2 where salt is lying well within the 50% explained variance circle.

This shows that relying only on a two-dimensional external preference mapping plot can leave important drivers of liking undetected.

10. Conclusions and suggestions

Using averages over assessors or raw data.

The average data will give a simpler solution to look at, but no information about individual differences across assessors in the panel. When choosing averages it is not possible to apply PC-ANOVA the way presented here for deciding on the number of components. If averaging is used, one should always do a proper check on the reliability of the individual assessors before averaging.

Standardisation

The calibration and training procedure should be considered and evaluated for making a decision on whether to standardise or not. The focus should be on the meaningfulness of relying on actual differences in variability of different attributes (possibly belonging to different sensory modalities) in the analysis. If these are not meaningful, one should standardise. This is an interesting aspect when comparing results from different panels. In such cases, the need for standardisation is stronger unless the training procedure is harmonised between the labs. If clearly non-significant variables are present, one should be careful about incorporating them in a standardised analysis.

Using all attributes or eliminating obvious overlap.

Eliminating highly correlated variables will in most cases have only a moderate effect on the interpretation. One should be careful about strong statements about what are the most

important sensory dimensions since this will depend on the number of attributes that represent it. A tool based on partial correlations is presented that can enhance insight into non-trivial overlap among attributes.

Validation of components

Leave-one-out Cross-validation is often not the best choice in sensory analysis when samples are unique and few.. In such cases an alternative is to use permutation testing.

Relating sensory QDA data to consumer liking data

In this case it is important to be aware that not all variables may be of interest. If obvious candidates exist, one should consider excluding the non-informative variables. On the other hand, there may be important attributes that are not so visible when considering only few principal components of sensory data. It is always recommended in such cases to compute a PCA model of consumer liking data to support the conclusions. Alternatively, one can take the latter as point of departure and regress sensory variables individually onto the PCA solution (internal preference mapping).

Acknowledgements.

We would like to thank Dr Nguyen for providing the yogurt data. The authors would like to thank for financial support from Research Council of Norway.

References

Bro, R., Qanari, E.M., Kiers; H.A.L. , Næs, T. and Frost, M.B. (2008). Multi-way models for sensory profiling data. J. Chemometrics, 22, 36-45..

Brockhoff, P.B., Schlich, P. and Skovgaard, I. (2015). Taking individual scaling differences into account by analyzing profile data with the Mixed Assessor Model. Food Quality and Preference 39, 156–166.

Cadoret, M., & Husson, F. (2013). Construction and evaluation of confidence ellipses applied at sensory data. *Food Quality and Preference*, 28, 106-115.

Clark, C. C. and Lawless, H. T. 1994. Limiting response alternatives in time–intensity scaling: An examination of the halo-dumping effect. Chemical Senses, 19, 583–594

Dahl. T. and Næs, T.(2004). Outlier and group detection in sensory analysis using hierarchical clustering and the Procrustes distance. Food Quali. Preference., 15 (3). 195-208.

Dahl, T., Tomic, O. Wold, J.P. and Næs, T. (2008). Some new tools for visualising multi-way sensory data. Food Quality and preference, 19

Dahl, T and Næs, T. (2009). Identifying outlying assessors in sensory profiling using fuzzy clustering and multi-block methodology. Food Quali. Peference, 20, 287-294.

Dijksterhuis, G. (1995). Assessing panel consonance, Food Qual. Preference, 6 (1), 7-14.

Endrizzi, I. Gasperi, F., Rødbotten, M and Næs, T. (2014). Interpretation, validation and segmentation of preference mapping models. Food Quality and Preference. 32, 198-209.

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S: Scaling. In Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS) Umetrics; 1999:213-225.

Escofier, B. and Pages, J. (1995).Multiple factor analysis, Comp. Stat. Data Analysis. 18, 121-150.

Gower, J. C. (1975). Generalized Procrustes analysis. Psychometrika, 40, 33–51.

Jansen, J., van der Hoefsloot, J., Greef, M., Timmerman, E., Westerhuis, J., & Smilde, A. K. (2005). ASCA: Analysis of multivariate data obtained from an experimental design. Journal of Chemometrics, 19(9), 469–481.

Jolliffe, I.T. (2010). Principal component analysis. Springer.

Lawless, H.T. and Heyman, H (2010). Sensory evaluation of food. Principles and practices, Springer Science and Business Media, New York.

Liland, K.H., Smilde, A. and Næs, T. (2018). Confidence ellipsoids for ASCA models based on multivariate regression theory. J. Chemometrics, 32. https://doi.org/10.1002/cem.2990

Luciano, G. and Næs, T.(2009). Interpreting sensory data by combining principal component analysis and analysis of variance. Food Quality and Preference, 20, 3, 167-175.

Martens, H. and Martens, M. (2001). Multivariate Analysis of Quality : An Introduction. John Wiley and sons, Chichester.

Nguyen, Q. C., Næs, T., Almøy, T and Varela, P. (2019). Portion size selection as related to product and consumer characteristics studied by PLS Path Modelling. Food Quality and Preference (in press)

711 Næs, T., Varela, P. and Berget, I. (2018). Analysing individual differences in sensory and
712 consumer science. Elsevier.
713
714 PanelCheck software: https://sourceforge.net/projects/sensorytool/, DOI:
715 10.5281/zenodo.10768

716 Romano,R., Brochoff, P.B., Hersleth, M., Tomic, O., Næs, T. (2008). Correcting for different
717 use of the scale and the need for further analysis of individual differences in sensory analysis.
718 Food Quality and Preference, 19, 197-209.

719 Schlich, P. (1996). Defining and validating assessor compromises about product distance and
720 attribute correlations. In Næs, T. and Risvik, E.(Eds.). Multivariate analysis of data in sensory
721 science. Elsevier, Amsterdam.

722 Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. J. Roy.
723 Stat. Soc., B. 111-133.

724 Stone, H., Bleibaum, R. and Thomas, H. (2012). Sensory evaluation in practivc.4th edition.
725 Academic Press.

726 Tomic, T., Nilsen, A., Martens, M. and Næs, T. (2007). Visualization of sensory profiling
727 data for performance monitoring. LTW, 40, 262-269.
728
729 Tomic,O., Luciano, G., Nilsen, A., Hyldig, G., Lorensen, K. and Næs, T. (2010). Analysing
730 sensory panel performance in a proficiency test using the PanelCheck software. Eur. Food
731 Res. Technology, 230, 3, 497-511.

732 Tomic, O., Forde, C., Delahunty, C. and Næs, T. (2013). Performance indices in descriptive
733 sensory analysis – A complimentary screening tool for assessor and panel performance. Food
734 Quality and Preference, 28, 122-133

Tucker, L.R., (1964). The extension of factor analysis to three-dimensional matrices. In N. Fredriksen and H. Gulliksen (eds). Contributions to mathematical psychology, Holt, Rinehart and Winston, NY.

Vitale, R., Westerhuis, J.A., Næs, T., Smilde, A.K., de Noord, O.E., Ferrer, A. (2017). Selecting the number of factors in Principal Component Analysis by permutation testing - Numerical and practical aspects. J. Chem. 31, 12, pages …

Wold, S. (1978). Cross-validatory estimation of the number of components in factors analysis and principal component models. Technometrics, 20, 397-406.

Data set	Number of samples	Number of attributes	Number of assessors
Yogurt	8	21	9
Olive oil	11 and 4	20	Only averages used
Bread	8	13	Only averages used

745

746 **Table 1.** Overview of QDA data sets used. For the olive oil data set also the small subset is
747 tested. For the bread data also consumer liking data for a number of consumers were available
748

Figure Captions

Figure 1a. Illustration of the setup in Equation 2. The D now represents the number of design variables (including product and assessor factors plus interactions).

Figure 1b Data structure for QDA presented as a three-way data set and an unfolded data set. The illustration is for simplicity only for 4 assessors. If replicates are present, the vertical dimension will be samples*replicates (I*R)

Figure 2, Yogurt data. Standardised PCA on consensus data, 20 significant attributes.

Figure 3. Yogurt data. Scores and loadings for the standardized PCA based on individual data, 20 significant attributes.

Figure 4a, Olive oil data. Full data set non-standardised

Figure 4b. Olive oil data. Full data set standardised

Figure 4c. Olive oil data. Reduced data set non-standardised

Figure 4d. Olive oil data. Reduced data set standardised

Figure 5, Olive oil data. Heat map of correlations between residuals for different attributes. Over all assessor in a). The other three, b), c) and d), represent three individual assessors.

Figure 6, Olive oil data. Non-standardised PCA, 14 significant attributes. The illustration in a) shows the curve obtained by the permutation method. The points represent the quantiles for each of the number of components. In b) is presented explained variance for fitting/calibration and leave-one-out cross-validation.

Figure 7. Olive oil data. Four samples, standardised PCA, 14 significant attributes. a) scores and b) correlation loadings, c) cross-validation, d) permutation testing.

Figure 8. Yogurt data. PCA-ANOVA results, standardised PCA, 20 significant attributes. a) multiple comparisons for products. Line indicates range of no significant differences. b) F-values for the product effect factor. The significance is indicated with colour as given in panel in the upper right corner.

Figure 9, Yogurt data. Standardised PCA. 20 significant attributes. Permutation test for PCA based on averages over assessors.

Figure 10, Bread data. Non-standardised PCA. Correlation loadings for external preference mapping . a) component 1 vs. Component 2. b) component 1 vs. Component 3.