**Interpretation, validation and segmentation of preference mapping models**

Isabella Endrizzia, Flavia Gasperia, Marit Rødbotten b and Tormod Næsb

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

b Nofima Mat AS, Osloveien 1, NO-1430 Ås, Norway and Dept. of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Fredriksberg Copenhagen, Denmark

**Abstract**

In this paper we discuss an extension to preference mapping of the method proposed in [Endrizzi, I., Menichelli, E., Johansen, S. B., Olsen, N. V., & Næs, T. (2011). Handling of individual differences in rating-based conjoint analysis. Food Quality and Preference, 22, 241-254 ] for accommodating both population averages and individual differences in the same model. The method, based on average estimates and residuals, is a combination of ANOVA, PCA and PLS-DA, which are well-known techniques that can be run in almost all statistical software packages. Main attention is given to the relation between the double centred residual matrix which highlights differences between consumers in their relative position as compared to the average consumer values and the standard centring in preference mapping. This approach has been found particularly useful for highlighting differences in preference pattern among the consumers. Furthermore, the interpretation and the segmentation, that is here taking place based on differences in acceptance pattern, are graphically oriented. In addition, some possible alternatives to the generally used validation method in PCA are suggested. The approach is then illustrated using two data-sets from consumer studies of apple and raspberry juice, showing that when individual differences are analysed by the present method, interesting results regarding individual differences in response pattern were detected.

Key words: Individual differences, preference mapping, ANOVA, PCA, PLS-DA, validation.

**1. Introduction**

Preference mapping (Chang & Caroll, 1969; Caroll, 1972) is an important and much used methodology for modeling, analysing and understanding consumer preferences and their relation to product characteristics. Because of its attractive properties, preference mapping has been used for a number of different purposes, for instance to identify sensory drivers of liking (Michon, O’Sullivan, Sheehan, Delahunty, & Kerry, 2010; Sinesio et al., 2010), to find the best product composition (Endrizzi, Pirretti, Calò, & Gasperi, 2009; Felberg, Deliza, Farah, Calado, & Donangelo, 2010) and as a method for product optimisation (Busing, Heiser, & Cleaver, 2010; Ares, Varela, Rado, & Giménez., 2011). The most used methods in the area are the internal and external linear preference mapping methods (McEwan, 1996; van Kleef, van Trijp, & Luning, 2006). The main advantage of these techniques is that they are simple to use and interpret and therefore provide also non-statisticians with useful information. In some cases one needs to extend the methodology for handling non-linear preference tendencies. This can be done through the ideal point modelling strategies based on polynomial regression (McEwan, 1996). A practical limitation to this type of methods, however, is that they require a larger set of samples for giving precise model estimates. A strategy based on serving the consumers different samples with subsequent fuzzy clustering was proposed in Johansen, Herseleth, & Næs (2010) for solving this problem. We refer to several authors (Carroll, 1972; Danzart, Sieffermann, & Delarue, 2004; Mullenet, Lovely, Threlfall, Morris, & Striegel, 2008; Rousseau, Ennis, & Rossi, 2012) for alternative ways of compensating for non-linearity problems. For an exhaustive comparison of methods and software programs to perform preference mapping we refer to Yenket, Chambers, and Adhikari (2011).

Although the main purpose of preference mapping is to understand individual differences and their relation to the product characteristics, in many cases one is also interested in understanding how these differences relate to consumer characteristics such as age, gender, attitudes and habits. This type of information can sometimes be essential for developing for instance successful marketing strategies. A number of different methods are developed for this purpose, both one step procedures (Martens et al., 2005, Vigneau, Endrizzi, & Qannari, 2011) and procedures based on first analysing the liking pattern and then relating this pattern to the external data (see e.g. Næs, Brockhoff & Tomic, 2010).

All these methods are useful, but there are still several aspects to consider in more detail. This paper is devoted to a discussion of a number of these aspects with focus on validation, segmentation and different centring of the preference data. In particular we will generalise the method proposed in Endrizzi, Menichelli, Johansen, Olsen, and Næs (2011) for conjoint data to incorporate also preference mapping. This is a method that accommodates both population averages and individual differences in the same analysis of variance approach. It is of particular interest to discuss how these methods differ when comes to segmentation. When concerns validation of preference mapping results, we will propose a couple of new approaches based on permutation and ANOVA testing which are particularly useful when the number of samples is very low, as is often the case in this type of studies. We will also highlight the importance of segmentation based on visual inspection of plots (Endrizzi et al., 2011) since it allows for segmentation according to interpretation and thus represents a more flexible approach than automatic procedures (Wajrock, Antille, Rytz, Pineau, & Hager, 2008). A major aspect here is also that one can seldom expect clearly separated consumer segments and different automatic clustering procedures can therefore easily end up with different proposals or proposals with little meaning. For illustrating the methodology we will use two data sets from consumer studies on apple juice and raspberry juice. In both cases, the general tendency of population liking has already been studied, but it will be shown that more information can be extracted handling individual differences as proposed here.

**2. Methods**

**2.1. Linear and ideal point preference mapping**

Linear preference mapping is based on relating sensory data to consumer liking data using a multivariate linear regression model, usually either PCR or PLS regression (McEwan, 1996, Martens, Esposito Vinzi, & Martens, 2007). For so-called internal preference mapping, the consumer data are treated as independent variables while for external mapping the sensory data are considered the independent ones. Both methods are used in practice and they have their own advantages and disadvantages as discussed in for instance by Næs et al.,2010. They both provide scores, sensory loadings and consumer loadings which are interpreted in the same way using scatter plots. Ideal point preference mapping (as discussed in McEwan, 1996) by the use of polynomial regression is strongly related to linear external mapping, the only differences being that quadratic and interactions terms between the principal components of the sensory data are added to the linear model. Both these methods will form the basis for the rest of the developments in this paper, with main focus on linear mapping.

**2.2. An alternative approach**

The single framework approach for conjoint analysis mentioned above (Endrizzi et al., 2011) is based on estimating the average population liking effects using a standard ANOVA model and considering the residuals for interpretation of the individual differences. In other words, the averages are calculated to represent the population structure and the residuals are calculated to represent the individual differences representing the deviations from the average effects. An advantage of this approach is that all results can be understood within the same framework model. For interpretation it is, however, important to note that the residuals obtained in this way are double centred as will be discussed below.

For our purpose, the appropriate ANOVA model can be written as

, i=1,…,I, j=1,…,J (1)

where yij is the (ij)th observation, μ is the general mean, αi is the main effect of the tested products, Cj is the random main effect of consumers and εij is the random error. Since each consumer tests each sample once, no interaction between consumer and product is possible. The information about individual differences and how they interact with the product effect is therefore to be found in the residuals and **only** there. The average liking for the different products will be found in the estimates of the ’s. With the product effect and the consumer main effect in the model, the residuals can be written as

 (2)

These residuals are double centred, i.e. they are mean centred across products and across consumers for each combination of i and j.

As for standard internal preference mapping these residuals can also be analysed by a PCA. In Endrizzi et al. (2011) it was illustrated how the scores and loadings plot from such a PCA can be used for visual interpretation and segmentation. If external consumer characteristics are available the consumer loadings for the PCA can be related to these values using for instance PLS regression. If useful segments are identified, one can use discriminant PLS instead. It is also possible to regress the sensory attributes onto the scores of this PCA in order to understand how the different dimensions relate to the sensory data.

**2.3. Interpretation of individual differences**

Although the residuals approach will here mainly be used for visual segmentation according to interpretation, it is worthwhile to discuss the relation between the PCA results of the raw data (identical to internal mapping) and the residuals.

Using only centring per consumer as done for standard internal mapping (like in MDPREF, see e.g. McEwan, 1996), the PCA results represent how the different consumers rate the relative differences between products. The concept “relative” here refers to the fact that the level of liking is removed due to the centring. For a meaningful interpretation of the PCA plot, one implicitly assumes that there is an underlying space of relative differences that all consumers have in common (scores), but one allows for the consumers to use this underlying space differently. This means that the consumers, which represent the loadings, may in principle be spread out all over the place in the plot. All this implies that the liking of the different **products** play the most important role in the interpretation. Note that regular MDPREF implicitly, since only averages for each consumer are subtracted, contains information about both average product effects and individual differences in liking.

If, however, the matrix is double centred as it is in the residual approach, the situation is different. In this case also the effect of the relative liking of the products is eliminated and the subsequent PCA has another interpretation. For each of the products (rows in the matrix), the sum is equal to 0 which means that the values in this row refer to how the different consumers score according to the average consumer for that product. The whole matrix is then instead focused on the relative differences of the consumers, relative to the average consumer value for each of the products. In others words, the results are concentrated on the consumers and how they relate to each other rather than to how they rate the relative differences between the products. Since the sum for each consumer is still 0 because of double centring, this analysis will then contrast consumers relative to which samples they give a score above or below the average consumer value. The consumers will always be above 0 for some samples and below 0 for others since the sum is equal to 0 both ways.

From a study of the individuals differences point of view (and thus segmentation), both the approaches are useful, but they tell us different things. Standard MDPREF focuses on liking of the products, while the other option focuses on relative differences of consumers after elimination of product effects. Since the residuals contain all possible information about the individual differences and **only** that (in addition to random noise) one can argue that the double centred approach may in some cases be even more important than the standard approach in for instance segmentation procedures. This aspect will be discussed below.

**2.4. Validation**

Validation of a model means testing its performance according to an a priori given set of test result specification (Esbensen, 2002). One of the main purposes of validation in preference mapping is also to determine the number of principal components that can safely be interpreted.

In this area where the focus is on interpretation on a given set of samples and not necessarily on predictive performance of the model for other samples, cross-validation (CV) of the samples is not always the natural choice. This is further underlined by the fact that the sample set is often based on a designed experiment with possibly unique samples in it. In addition, the sample set may be very small. Therefore, it may be more natural to proceed in the direction of internal validation related to the actual samples at hand instead of some type of predictive validation. There are a number of possible approaches to this which will be discussed next. For internal mapping one of them is to use a permutation test for testing the significance of the principal components extracted from the consumer liking data. Another approach is to validate the consumers instead of the samples. The latter means testing the fit of the different consumers to the PCA model determined by the rest of the consumers, but this approach may have strong limitations as discussed below. A third possibility for external mapping is to regress the individual consumer onto the sensory space and determine significance of each by regular F-testing.

A general argument against CV for PCA performed by eliminating rows in the data set is that the fitting of an object to the model fitted by the rest will in most cases lead to an increased explained variance. The reason for this is that, as opposed to CV used in regression, each predictive fitting will lead to a decreased residual sum of squares when a new component is incorporated. Some of this can be compensated for by correcting for the number of degrees of freedom in the fitting procedure (Martens & Næs,1989), but only in rear cases will there be an increased explained variance followed by a decrease, meaning that determining the number of components may be difficult. This effect is particularly strong when the vector of fitted values is low-dimensional as it is when fitting a new consumer to the model fitted by the rest. In other words, using CV for the consumers as mentioned above, may give an overoptimistic view on the validity of a component. This will be illustrated below.

*2.4.1. Cross-validating consumers rather than samples*

Since this procedure tests the fit of the consumers to the general model determined by the rest of the consumers, it can also be useful for identifying those consumers who have a very low explained variance and thus do not follow the general pattern defined by the rest of the consumers. A possible and simple criterion to use for this is to compare with for instance 50% explained variance after two components or the average explained variance (over consumers) for the same number of components. The same basic idea is used in a PLS context by Tenenhaus, Pagès, Ambroisine and Guinot (2005) for eliminating from the analysis those “individuals” with a low predictive performance.

*2.4.2. Permutation testing for internal preference mapping*

Permutation testing is a non-parametric tool that allows for evaluating statistical significance for a null-hypothesis by repeatedly randomising the original data-set (Good, 2000). Note that, since all permutations have the same probability of occurring, a large set of total possible permutations (e.g. 300 or more) should produce a representative sampling distribution (Xiong, Blot, Meullenet, & Dessirier, 2008). In the literature there are a number of methods available for permutation testing in PCA (Wakeling, Raats, & MacFie, 1992; Landgrebe, Wurst, & Welzl, 2002; Peres-Neto, Jackson, & Somers, 2003; Linting, van Os, & Meulman, 2011), but as far as we know, there is little available for testing the significance of each additional component which is the focus here. The procedure used in this paper is described in detail in the Appendix and as far as we know this is new. Again the interpretation is in terms of internal validity and different from regular CV for the samples.

*2.4.3. Validation for external preference mapping using regression F-values*

The basis for this procedure is to regress liking values for the consumers onto the principal components of the sensory data (PCR). In this type of validation one has a choice with respect to which model to use. In the procedure proposed here, we will use both linear and polynomial models. First we check whether the polynomial model is significant (Næs, Isaksson, Fearn, & Davies, 2002). If it is, we propose to test the effect of the non-linear part. If that is significant, the model is determined to be non-linear. If not, it is determined to be linear. For the non-significant non-linear models, one uses a linear model and tests for significance and then splits them into linear and non-significant ones.

Note that in a certain sense this approach is similar to eliminating consumers based on explained variance as was discussed above. Similarly to that approach, the consumers corresponding to the non-significant models could be eliminated from further PCA analysis since these consumers do not fit the general model structure and should therefore be considered separately. There could be several reasons for this lack of fit, some consumers may for instance have problems discriminating between the samples, or they could represent a completely different model structure. Anyway, analysing them together with the rest of the consumer is questionable. The consumers with a non-linear structure should in principal also be considered separately.

If two principal components are used, a second-degree polynomial regression model can be written as

 (3)

where the *b*s are the regression coefficients to be determined and *f* is the random error term. In consumer studies the number of products is often so low that estimating so many parameters is not possible. One way to solve this problem is to use a simplified model, for instance the following where the two PCs were combined in one non-linear term and the interaction was eliminated:

 (4)

This model will be used for the polynomial regression models analysed below.

**2.5. Visual segmentation vs. automatic segmentation procedures.**

Statistical clustering procedures can if wanted be used for segmentation purposes in consumer science. In the literature there are a number of different methods that have been used for this purpose (see among others Kaufman & Rousseeuw, 1990; Vigneau, Qannari, Punter, & Knoops, 2001; Wajrock et al., 2008; Everitt, Landau, Leese, & Stahl, 2011). Although useful in many cases, a possible problem with these approaches in consumer science is that there is seldom a clear separation of consumers into clusters. In other words, in many cases, there is only a continuum of individual differences with no clear separation point between them. Using an automatic clustering procedure which is developed for something else can therefore be questionable. The results may certainly be useful and meaningful, but there is no guarantee that the groups are identified according to a meaningful interpretation. The clustering may also depend heavily on which method that is used, which will be shown below.

Another and sometimes better possibility is then to use visual inspection of the individual differences results. An important possibility that was discussed in Endrizzi et al. (2011) is to segment according to interpretation of the PCA results of the individual differences. This can be done using a regular preference mapping approach or using the residuals approach discussed above. In the PCA results there will be one plot for the individual consumers and one for the products. The idea is to use the interpretation of the components according to which products they represent, and then make a segmentation according to what one is interested in. In the results section this type of procedure will be tested and compared to more automatic procedures.

For the illustration, we will concentrate on the K-means method, the most popular partitioning clustering algorithm, and a few so-called hierarchical methods (HCA). We will use two of these, complete linkage and Ward linkage. For a more extensive clustering method comparison still in the framework of preference mapping we refer to Wajrock et al. (2008). It is worth mentioning that segmentation by one of these automatic methods will give the same results for both consumer centred data (the type of data used for regular MDPREF) and double centred data. The reason for this is that one of them is obtained by just subtracting the average product profile from the other (mean centring).

**3. Data sets used.**

**3.1. Apple juice data**

This dataset has previously been used for investigating consumer liking of apple juices produced with different levels of sugar and acid (Rødbotten et al., 2009). The design used was a full factorial design with two factors of interest: degree of sugar at three levels (low, medium and high) and degree of acid at two levels (low and high). The prepared six juice combinations were first evaluated by a sensory panel on a list of 14 descriptors, and then by a panel of 125 consumers who were asked to rate their degree of liking on a 7-point hedonic scale for each juice. After the consumers had rated the samples, they were asked to fill in a questionnaire related to health attitudes and habits in drinking apple juice. Identification of the samples in the plot, the first letter in the symbol refers to sugar level (H-High, M-Medium, L-Low) and the last letter refers to level of acid (H-High, L-Low).

**3.2. Raspberry juice data**

The primary aim of this study was to investigate the acceptability of 25 juices created by mixing one of the five berry fruits under study with five different base juice variants (Endrizzi et al., 2009). Seventy-two consumers were involved in a series of five central location tests, each of them focused on one of the five berry fruits investigated. Here, for illustrative purposes, only data from juices based on raspberry were considered. In the test session consumers were first asked to rank the mixes according to their overall liking, subsequently to rate their appreciation on a 9-point scale and then fill a questionnaire about their habits and opinions on juice consumption. In this case, sensory data are not available, but chemical analyses were carried out for providing information about compositional parameters and antioxidant capacity. In the external mapping discussed below, these chemical data will be used instead of sensory data which are normally used in preference mapping. In the plot, sample codes are assigned according to the ingredient mixed with raspberry (A: apple, BO: blood orange, O: orange, P: pineapple and PG: pomegranate).

**4. Results and Discussion**

Calculations in this paper were conducted in Statistica 9.1 (Statsoft Inc., 1984-2010), The Unscrambler®X 10.2 (CAMO Software AS, 2009-2012) and Matlab 7.9 (The MathWorks, 2009). Except for the permutation test, all the proposed method can be undertaken using commercial software without the need for computer programming.

**4.1. Results for apple juice data**

In order to clarify and visualise the differences between analysing original liking data and the ANOVA residuals, the PCA loading and score plots for both analyses were computed (see Figure 1 and Figure 2).

The explained variance for the two first components based on the original data was 74%. The first component is strongly related to the level of sugar in the products: negative values of PC1 correspond to low level of sugar (product names starting with L) and positive values to high level of sugar (product names starting with H). The second component separates juices according to level of acid: negative values of PC2 correspond to low level of acid (product names ending with L) and positive values of PC2 correspond to high level of acid (names ending with H). As can be seen, almost all consumer prefer the sweeter juices, while the population is split according to which acid level is preferred.

The results from the ANOVA mixed model (according to model (1) above) are reported in Table 1. As can be seen, the main factors are both highly significant. Products have the strongest effect and the product averages show that there are three groups according to sugar level (HSD Tukey’s test), corresponding very well with the PCA results above. No clear separation can be seen here with respect to average differences between the acid levels.

For the residual data (Figure 2), the explained variance for the two first components was 60%. The first and most dominant component is in this case strongly related to the acidity of the products tested. On the left side of the plot we find consumers who prefer juices with low level of acidity (product names ending with L) whereas those who appreciate a higher acidity as compared to the average are located at the right side (names ending with H). The second component separates products with high level of sugar (names starting with H), positioned in the lower part of the plot, from juices with medium or low sugar content (names starting with M of L) related to positive values of PC2.

*4.1.1. Comparing internal preference mapping and the residual approach*

As can be noted, the first plot (Figure 1) which focuses on products reveals that sugar and therefore sweetness is the most important driver of liking for the products, while the second plot (Figure 2) clearly shows that when coming to the individual differences after subtracting of the mean liking values, the acidity is the most important for distinguishing between individuals. In other words, the two plots supplement each other for the interpretation of the data.

There are consumers in all quadrants in the loadings plot for the residuals approach, which means that it lends itself very naturally to a visual segmentation. It should be noted that also the Figure 1 can be segmented according to a visual segmentation, but in that case, the most natural way of doing it would probably be just to separate according to acid level. In this sense, the residuals plots provide in this case a more obvious and “richer” opportunity to segment according to interpretation.

*4.1.2. Visual segmentation*

The segmentation chosen for visualisation (see Figure 2) distinguishes between consumers preferring different categories of samples represented by the scores. For instance, the segment G1 consists of consumers with the strongest preference for high acidity as compared to the rest of the consumers. Likewise, G3 goes more in the direction of low acidity. The groups G2 and G4 distinguish between different preference for sweetness. It is important to note that this is only one of several possibilities for segmentation depending on which type of interpretation one is interest in. One could for instance only focus on preference for sweetness , thus think of a segmentation only along the first component distinguishing between those who prefer the highest sugar content and the rest out.

In order to visualise and verify these interpretations, the sample average liking values and their 95% confidence intervals for each of the groups are presented in Figure 3. There are a number of important things to note here. First of all, G4 consists of consumers who clearly prefer the highest sugar level and dislike the rest. They do not distinguish between low and high acid level except for high sugar level samples. G3 is a segment which clearly likes samples with low level of acidity and has a slight increase in liking when the sugar level increases. G1 is more or less the opposite. G2 is a segment which represents small changes in liking for all samples except a small increase when sugar goes from the lowest to the two higher levels. All these interpretations correspond well with the positioning of the scores as compared to the segments in the loadings plot, but give a more explicit and clear interpretation of the various segments.

*4.1.3. Relating groups to consumer information*

For illustration we then tried to characterise the four groups in terms of collected consumer attitude data. To do this, PLS-DA was used on consumer attitude data where dummy group variables were used as Y. Here we report the comparison between group G1 and G3 which are very different for their liking of acidity.

First we considered the model with all the attitude variables involved: Only three of them showed a significant relation to the groups according to the jack-knife method (Martens & Martens, 2000). A new analysis with only the significant variables was then conducted (Figure 4). It can be noted that the interpretation of those variable does not differ in the two models (i.e. the full and this one). Consumers in G1 are consumers who think it is important to eat healthy food and prefer juices with no added sugar beyond that already contained in the raw material. In group G3 the consumers think that it is more important to avoid sugar for elderly people than for children. These results make sense in terms of the interpretation of the acceptance pattern given above, i.e. those who like the juices with higher acidity level (G1) show a more positive attitude for healthy aspects. For instance, consumers who prefer juices with no extra sugar are probably more used to acid taste. The explained variances were 72% for the X and 8% for two first Y components. For cross-validation of the samples (see below), only the first component was important for both X- (30%) and Y-variance (6%).

*4.1.4. Validation based on cross-validation*

In table 2a, cumulated explained variance for each principal component using validation on samples and on consumers for the apple juice data set are reported. The explained variances from CV of the samples are lower than for fitting, but it clearly indicates two components as significant (only limited increase after), which corresponds well with the two factor design. From this perspective, the standard CV seems to work quite well, which is due to the fact that the design is very simple and the samples have a lot in common. For the CV of the consumers, the CV values indicate a higher number of components which also emphasises that it has another interpretation. But as we can se, the explained variances always increase, which supports the warnings in Section 2.4 above.

The 50% test (based on CV) for eliminating consumers suggested in section 2.4 gave that 39% of the consumers had an explained variance value higher than 50% after two components. If we consider as critical value the average percentage for all consumers after two components (31%) 54% of consumers are well explained by the model. We then did a new preference mapping with these consumer eliminated, but the results were very comparable to the general results with all consumers in there. This corresponds well with previous experience (Helgesen, Solheim, & Næs, 1997) that even though one may argue that this is the correct way of doing it, the overall conclusions do not change much.

*4.1.5. Validation using the permutation test*

On the same data set, the permutation test was run with B=300 in order to evaluate the significance of each component. In Figure 5a, the comparison of observed explained variance with that obtained in the permutation test are depicted. Note that the measure used for testing significance of a new component is explained variance relative to the sum of the variances for the remaining components (see also Appendix). Because the observed values of variance of the first two components are larger than the 95th percentile, we have to conclude that these two components are clearly significant. The corresponding *P*-values for each component were also calculated and they were *p*1= 0.010, *p*2= 0.010, *p*3= 0.218, *p*4=0.802 confirming the conclusions given by the graph. Note then in this case, 5 components is the maximum with 100% explained variance for all possible validation methods. Here 5 components (100%) is included in the plot for comparison.

*4.1.6. Validation using the F-test procedure*

The procedure of using linear and polynomial models discussed for external mapping above was also used in this case. Choosing 0.1 as critical value, three groups of consumers were identified: 14 non-linear, 49 linear and 62 non significant ones. No relation between these groups and those obtained with dividing the preference map with 4 segmentation sectors in the residual approach were found. It was found that 72% of the non significant consumers here correspond to the non-significant ones in the 50% approach above, and 54% in the average% approach (see Section 2.4.1).

*4.1.7. Comparing automatic segmentation techniques*

The K-means and HCA with two different linkage criteria (complete and Ward linkage,) were used for the original liking. Studying the dendrogram, three clusters were identified as a natural choice (Ward’s method, Euclidean distance). The cross-tabulation of group assignment for the different segmentation techniques is reported in Table 3. As can be seen for Ward’s method and K-means clustering the G1 and G3 are reversed, but this no effect on the results. Apart from this, the results are quite comparable; only a small number of consumers are exchanged. The grouping obtained with HCA using complete linkage is the most different with two clear groups, one of them being a combination of Ward’s G1 and G2. In Figure 6 the clusters obtained with complete linkage and K-means clustering are visually compared. K-means clustering with four groups is also depicted and compared to visual clustering. As can be seen, the results are very different. As shown in the example, clusters obtained by an automatic method can give quite different results, indicating the advantage of visual clustering as done above.

**4.2. Results for Raspberry data**

The results of the ANOVA mixed model are presented in Table 4. The product factor is highly significant whereas there is no effect of consumer. Comparing the MS’s, the product effect accounts for as much as 77% of the variation. From this ANOVA we computed the double centred residuals, and put them into a matrix with the samples as rows and consumers as columns (5 x 72). A PCA was run and loading and score plots were interpreted (Figure 7). For comparison we did the same for original data, which correspond to regular internal preference mapping (Figure 8). The two first principal components of the residuals explain 79% of the total variability while for the raw data they account for 72% of the variation.

*4.2.1. Comparing internal preference mapping and the residual approach*

As can be seen from Figure 8 most consumer lie on the upper right side of the loadings plot. The linear preference direction is indicated by an arrow in Figure 8a. The products BO, P and O are most liked, some prefer product PG while almost nobody seems to prefer product A. The PCA on the residuals (Figure 7) emphasises instead the differences between each consumer and the average consumer for each product. In this case the first component primarily separates the juice based on pomegranate (PG) from those based on orange (BO, O) while the second component separates the product based on apple (A) from PG, BO and O with P in the middle. As can be seen, the position of P is quite different as compared to the rest in the two plots. It can also be seen that there are consumers in all quadrants which is natural since consumers are compared to the average consumer for each product (in Figure 7).

Although it can be useful to interpret the PCA plots from the residuals directly, the main advantage lies in the fact that due to double centring there are products and consumers in all quadrants and this usually provides quite obvious ideas for segmentation. In this case, there are several options, but we have decided to split according to the lines presented in Figure 7. This means that we would like to highlight those consumers with a higher preference value than average (over all consumers) for products BO and O, for PG and for A and P. For the internal mapping based on raw data, the plot contains both products averages and individual differences and how to make a visual interpretable segmentation is here less obvious. In this case, one would probably go for a segmentation based on spitting between the right sections (quadrant 1 and 4) and upper left (quadrant 3). Few other ideas turn out to be natural in this case.

The profiles of the three groups are reported in Figure 9. The G1 (n=28) and G2 (n=28) are in some aspects opposite to each other: consumers in the first group dislike (or like less) the juice mix based on pomegranate whereas those of the second group like it. Consumers in group G3 (n=16) like juices based on pineapple and apple, which are the most and the least preferred juice by the rest of the consumers respectively. Interestingly, although the internal map based on raw data indicates that nobody likes the A juices, the G3 plot actually shows that there is a segment of consumers that give a high score for A. This shows that clustering according to residuals can provide new insight that is not easily available using the standard method. Note that all these results also show that even though there is a general tendency as reported by the main population effects above, there are large individual differences, and some of the consumers even rate the products in opposite order as compared to the average.

*4.2.2. Relating groups to consumer information*

The next step is to relate the three groups to consumer’ characteristics. For illustration here we concentrated on questionnaire data related to consumer habits and opinions on juice consumption. Categorical variables were transformed into dummy variables and weighted in order to have the same contribution in the model whereas quantitative variables were centred and standardised. The consumer characteristic variables were used as X-variables in a PLS-DA model with group identification as Y. Here the model only containing the significant variables (according to the jack-knife) is reported in the illustration (Figure 10). Consumers in G1, compared to those belonging to the other two groups, are regular drinkers and lovers of freshly squeezed and nutritious juices, in particular that of orange. On the contrary, in G2 consumers think that it is not important for a juice to be nutritious whereas those in G3 do not like orange juice. These results make sense in terms of the interpretation of the acceptance pattern given above: Those who claim to prefer orange juice confirmed this statement assigning higher positive score to juice mixes based on orange and vice versa. The explained variances were 41% for the X and 26% for the Y two first components (both suggested as significant according to cross-validated samples based on Y-variance (15%)).

*4.2.3. Validation based on cross-validation*

In table 2b, explained variances for each principal component using validation on samples and on consumers for the raspberry data set are reported. In this case, where the samples are much more different than in the previous case, the standard CV does not work so well. The variances explained by the CV are much lower than those for calibration and there is no clear indication on the number of components to use. Validating on the consumers again gives a continuous increase in explained variance supporting the warning in Section 2.4.

In this case only 10% of the consumers had an explained variance higher than 50% after 3 components. If we use the average % explained variance after 3 components as many as 60% have a higher value. Eliminating the “non-significant” consumers in the average % approach we noted an increase in the explained variance especially for the first component (from 46% to 60%) whereas for the second component there is a decrease (from 26% to 20%). Nevertheless, the samples separation according to the first two components in the product map is similar with BO, O and P separated from PG and A.

*4.2.4. Validation using the permutation test*

In Figure 5b, the comparison of observed explained variance with that obtained in the permutation test is depicted. Because the observed value of variance of the first three components is larger than the 95th percentile, we have to conclude that the first three components are significant. *P*-values for each component were also calculated (*p*1= 0.010, *p*2= 0.020, *p*3= 0.010) confirming the conclusions given by the graph. The plot of the first vs. the third component also indicates a structure related to component 3. Again the last component will always be 100% due to the fact that 4 components is the maximum.

*4.2.5. Validation using the F-test procedure*

Following the procedure described in Section 2.4.3. for external mapping the following three groups of consumers were obtained: 4 non-linear, 6 linear and 62 non significant (86%), 36% of non significant consumers more than those obtained from the first example. This low number of significant models is probably due to the fact that the relationship between preference and basic chemical composition is usually weaker than the link with sensory attributes. Also in this case, no relation between these consumer groups and those obtained with dividing the preference map with visual segmentation in the residual approach was found. It was found the 89% of the non-significant consumers in the F-test validation correspond to the “non-significant” ones in the 50% approach, and the 40% to the “non-significant” ones in the average% approach both described in Section 2.4.1. above.

Comparing the external preference map results from a PCR of preference data for all consumers onto chemical parameters and a similar PCR without the 62 non-significant, the explained Y-variance of the first two PCs increased from 49% to 78%. It is, however, questionable to interpret result from a preference mapping on just 10 consumers.

*4.2.6. Comparing automatic segmentation techniques*

As above the number of clusters to consider was based on dendrogram results from HCA (Ward method, Euclidean distance). In this case, two clusters turned out to be a natural choice. Then the K-means and HCA with complete and Ward’s linkage are used to segment consumers into two segments. The clustering results are quite different as illustrated in Table 5, where Ward’s and Complete grouping are the most different. In Figure 11, Complete and K-means clustering are compared and as can be seen, the results are quite different. Again the results indicate the advantage of using an interpretation driven segmentation.

**5. Conclusions**

This paper has been devoted to three different aspects of preference mapping: validation, segmentation, and different centring of the data.

We have discussed a number of alternatives to the commonly used cross-validation of samples for validating PCA results. The alternative methods considered are validating the consumers, significance of regression models based on sensory data and a permutation test. The authors recommend the use of the proposed permutation test that calculates the significance of each component providing the number of components that can “safely” be interpreted. The test results are easy to interpret thanks also to the graphical output. A particular advantage of using the test appears when the data are complex as they are in the second case reported here.

Both automatic segmentation procedures based on cluster analysis and segmentation based on visual inspection and interpretation of PCA plots are considered and discussed. It is shown that different automatic procedures can give quite different results due to the fact that there is no natural splitting of consumers into subgroups only continuum of different liking patterns. Segmentation based on interpretation is therefore recommended. In this way one can also more easily determine segments according to prior knowledge and focus of the study. The relation between segments and external consumer information were calculated in both examples and the results support the validity and interpretation of the clusters

Regards which plot to use for interpretation, the residuals approach is compared to classical internal preference mapping: first of all the corresponding PCA plots tell us different things and are therefore useful to different purposes. While the standard procedure focuses on differences in liking of the samples, the residuals approach focus more on the individual differences between the consumers and the average consumer. For this reason it seems that it offers some better possibilities for doing interpretation based segmentation. Both can be used for the purpose, but a more detailed view of the individual differences appear when using the residuals approach.

**Acknowledgements**

This work was partially supported by the Autonomous Province of Trento, Italy (AP 2010/2011). The contribution of prof. Tormod Næs was supported by the ConsumerCheck project (supported by the National Research Council of Norway and Norwegian food industry) and the Food Choice project (supported by The Agricultural Food Research Foundation).

**References**

Ares, G., Varela, P., Rado, G., & Giménez, A. (2011). Identifying ideal products using three different consumer profiling methodologies. Comparison with external preference mapping. *Food Quality and Preference*, 22, 581-591.

Busing, E. M. T., Heiser, W. J., & Cleaver, G. (2010). Restricted unfolding: Preference analysis with optimal transformations of preferences and attributes. *Food Quality and Preference*, 21, 82-92.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.). *Multidimensional scaling: Theory and application in the behavioural sciences* (Vol. 1, pp 105-155). New York: Seminar Press.

Chang, J. J., & Carroll, J. D. (1969). *How to use MDPREF, a computer program for multidimensional analysis of preference data* (Tech. Rep.). Murray Hill, Nj: Bell Telephone Laboratories.

Danzart, M., Sieffermann, J. M., & Delarue, J. (2004). New developments in preference mapping techniques: inding out a consumer optimal product, its sensory profile and the key sensory attributes. In *6th Sensometric Meeting*, Davis, CA USA.

Endrizzi, I., Menichelli, E., Johansen, S. B., Olsen, N. V., & Næs, T. (2011). Handling of individual differences in rating-based conjoint analysis. *Food Quality and Preference*, 22, 241-254.

Endrizzi, I., Pirretti, G., Calò, D. G., & Gasperi, F. (2009). A consumer study of fresh juices containing berry fruits. *Journal of Science of Food and Agriculture*, 89, 1227-1235.

Esbensen, K. H. (2002). *Multivariate data analysis - in practice: An introduction to multivariate data analysis and experimental design* (5th ed.). Oslo: Camo Process AS.

Everitt, B.S., Landau, M., Leese, S., and Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester, UK: John Wiley & Sons Ltd.

Felberg, I., Deliza, R., Farah, A., Calado, E., & Donangelo, C. M. (2010). Formulation of a soy-coffee beverage by response surface methodology and internal preference mapping. *Journal of Sensory Studies*, 25, 226-242.

Good, P.(2000). *Permutation tests: A practical guide to resempling methods for testing hypothesis* (2nd ed.). New York: Springer.

Helgesen, H., Solheim, R., & Næs, T. (1997). Consumer preference mapping of dry fermented lamb sausages. *Food Quality and Preference*, 8, 97-109.

Johansen, S. B., Herseleth, M., & Næs, T. (2010). A new approach to product set selection and segmentation in preference mapping. *Food Quality and Preference*, 21, 188-196.

Kaufman, L., & Rousseeuw, P. J. (1990).  Finding groups in data: An introduction to cluster analysis. New york: John Wiley & Sons Ltd.

Landgrebe, J., Wurst, W., & Welzl, G. (2002). Permutation-validated principal components analysis of microarray data. *Genome Biology*, 3, 0019.

Linting, M., van Os, B.J., & Meulman, J.J. (2011). Statistical significance of the contribution of variables to the PCA solution: an alternative permutation strategy. *Psychometrika*, 76, 440-460.

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., & Næs, T. (1989). *Multivariate Calibration*. Chichester, UK: John Wiley & Sons Ltd.

Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L.H., Høy, M., Westad, F., Thybo, A.

& Martens, M. (2005). Regression of a matrix on descriptors of both its rows and of its

columns via latent variables: L-PLSR. *Computational Statistics and Data Analysis*, 48,103–

123.

Martens, M., Esposito Vinzi, V., & Martens, H. (2007). The use of partial least squares methods in nw food product development. In H. Mafie (Ed.). *Consumer-led food product development* (pp 492-523). Cambrige: Woodhead Publishing Limited.

McEwan, J. A. (1996). Acceptance mapping for product optimization. In T. Næs, & E.Risvik

(Eds.), *Multivariate Analysis of Data in Sensory Science* (vol. 16 – Data handling in science

and technology, pp. 71–102). Elsevier Science B.V.

Michon, C., O’Sullivan, M. G., Sheehan, E., Delahunty, C. M., & Kerry, J. P. (2010). Investigation of the influence of age, gender and consumption habits on the liking of jam-filled cakes. *Food Quality and Preference*, 21, 553-561.

Mullenet, J.-F., Lovely, C., Threlfall, R., Morris, J. R., & Striegel, R. K. (2008). An ideal point density plot method for determining an optimal sensory profile for muscadin grape juice. *Food Quality and Preference*, 19, 210-219.

Næs, T., Brockhoff, P.B., & Tomic, O. (2010). *Statistics for sensory and consumer science*. Chichester, UK: John Wiley and Sons.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user friendly guide to Multivariate Calibration and classification*. Chichester, UK: NIR Publications.

Peres-Neto, P.R., Jackson, D.A., & Somers, K.M. (2003). Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology*, 84, 2347-2363.

Rousseau, B., Ennis, D. M., & Rossi, F. (2012). Internal preference mapping and the issue of satiety. *Food Quality and Preference*, 24, 67-74.

Rødbotten, M., Martinsen, B. K., Borge, G. I., Mortvedt, H. S.,Knutsen, S. H., Lea, P., & Næs, T. (2009). A cross-cultural study of preference for apple juice with different sugar acid contents. *Food Quality and Preference*, 20, 277-284.

Sinesio, F., Cammareri, M., Moneta, E., Navez, B., Peparaio, M., Causse, M., & Grandillo, S. (2010). Sensory quality of fresh French and Ducth market tomatoes: A preference mappingstudy with Italian consumers. *Journal of Food Science*, 75, S55-S67.

Tenenhaus, M., Pagès, P., Ambroisine, L., & Guinot, C. (2005). PLS methodology to study relationships between hedonic judgements and product characteristics. *Food Quality and Preference,* 16, 315-325.

van Kleef, E., van Trijp, H. C. M., & Luning, P. (2006). Internal versus external preference analysis: An exploratory study on end-user evaluation. *Food Quality and Preference*, 17, 387-399.

Vigneau, E., Endrizzi, I., & Qannari, E. M. (2011). Finding and explaining clusters of consumers using the CLV approach. . *Food Quality and Preference*, 22, 705-713.

Vigneau, E., Qannari, E. M., Punter, P.H., & Knoops, S. (2001). Segmentation of a panel of consumrs using clustering of variables around latent directions of preference. *Food Quality and Preference*, 12, 359-363

Wajrock, S., Antille, N., Rytz, A., Pineau, N., & Hager, C. (2008). Partitioning methods outperform hierarchical methods for clustering consumers in preference mapping. *Food Quality and Preference*, 19, 662-669.

Wakeling, I. N., Raats, M. M., & MacFie, H. J. H. (1992). A new significance test for consensus in generalized procrustes analysis. *Journal of Sensory Studies*, 7, 91-96.

Yenket, R., Chambers E. IV, & Adhikari, K. (2011). A comparison of seven preference mapping techniques using four software programs. *Journal of Sensory Studies*, 26,135-150.

Xiong, R., Blot, K., Meullenet, J. F., & Dessirier, J. M. (2008). Permutation tests for Generalized Procrustes analysis. *Food Quality and Preference*, 19, 146-155.

**Appendix**

The measure used for testing significance of a new component is explained variance relative to the ums of the variances among the remaining components. For instance, for component 2, the criterion used is the eigenvalue for component 2 as compared to the sum of the eigenvalues for components 2, 3, 4 etc. The permutation for the first component is simple since this relates to simple permutations of the original data set. For the rest of the components, however, it is more complex since residuals from previous components lie in a subspace orthogonal to previous components. This is here solved by orthogonalising permuted residuals with respect to both scores and loadings already estimated. This, however, changes the sum of variances for the permuted values and therefore the relative measure of performance just described must be used.

1. A principal components analysis on the preference data table Y (nxp) is performed and for each component, the percentage of explained variance after subtracting the amount of preceding components is recorded;
2. The following procedure is performed B times (for instance 300 times or more):
   1. A random permutation of the values located in each column of Y (consumers) is undertaken, independently of the other columns. Let Yperm be the resulting matrix. The variances of the variables in this matrix are equal to those of variables of the original Y. Nevertheless, all information about samples is lost and thus the correlation among the Yperms variables should be very small for each permutation.
   2. A PCA on Yperm is undertaken and the percentage of explained variance of the first component is recorded.
3. For the remaining components, the following procedure is performed n-2 times:
   1. Residuals from preceding principal component is calculated.
   2. A PCA on the residuals is undertaken.
   3. A random permutation of the residuals located in each column, independently of the other columns is undertaken and performed B times.
   4. The orthogonal projection of permutated residuals is calculated. Projections are made in both directions relative to the components already computed.
   5. A PCA on orthogonal permutated residuals is undertaken and the percentage of explained variance of the relative component is recorded.
4. The median, the 5th and the 95th percentile of the B values of explained variance from permutation are calculated and plotted. If the observed value of variance of a given component is larger of the 95th percentile, we can conclude that the first component is significant. Whether the observed value of variance is within the interval of confidence for the permutated data or is less than the 5th percentile, the conclusion is that first component is not significant. In order to facilitate the interpretation, it is also possible to calculate the *p*-value for each component as (*m* +1) / (B+1), where *m* typically represents the number of times that the explained variance observed for that component is larger than the permutated one. Then, we can reject the hypothesis that the observed explained variance for a specific component is equal to that due to noise/by chance if the p-value is less than α.

**Tables**

**Table 1** – Results from the ANOVA table of apple juice data: P-values for main effects (0.000 means a P-value ≤ 0.0001), averages and standard errors for the product effect are reported. The first letter in the product names indicate sugar level and the second letter the acid level (L: low, M: medium, and H: high). Different letters indicate different HSD Tukey’s test homogenous groups.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Sum Sq.** | **d.f.** | **Mean Sq.** | **F** | **Prob>F** |
| Products | 603.26 | 5 | 120.652 | 73.31 | 0.000 |
| Consumers | 380.01 | 124 | 3.065 | 1.86 | 0.001 |
| Error | 1020.41 | 620 | 1.646 |  |  |
| **Product averages (standard error)** | | | | | |
| **LL** | **LH** | **ML** | **MH** | **HL** | **HH** |
| 2.516a (0.117) | 2.389 a (0.111) | 3.560 b (0.114) | 3.560 b (0.137) | 4.480 c (0.132) | 4.792 c (0.123) |

**Table 2** – Cumulated explained variance in calibration and in validation using samples and consumers of apple juice (a) and raspberry data set (b). Note that 5 and 4 components are the maximum number of components in the two cases respectively, giving an exact explained variance equal to 100%. They are therefore omitted from the table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Validating samples** | | **Validating consumers** | |
| **a) Apple data** | **Cal.** | **Val.** | **Cal.** | **Val.** |
| PC\_0 | 0 | 0 | 0 | 0 |
| PC\_1 | 52.760 | 26.516 | 34.368 | 17.209 |
| PC\_2 | 74.118 | 35.626 | 60.271 | 35.084 |
| PC\_3 | 85.335 | 37.342 | 77.108 | 46.430 |
| PC\_4 | 93.226 | 37.089 | 89.434 | 56.626 |
| **b) Rasp data** |  |  |  |  |
| PC\_0 | 0 | 0 | 0 | 0 |
| PC\_1 | 45.890 | 7.630626 | 48.293 | 32.163 |
| PC\_2 | 72.25721 | 6.530319 | 78.996 | 62.098 |
| PC\_3 | 93.02387 | 16.2653 | 91.317 | 76.095 |

**Table 3** – Cross tabulation of group assignment of automatic segmentation methods for apple juice data set. In bolt count higher than 10 is highlighted.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **HCA Complete** | | | **K-means** | | |  |
| **HCA Ward’s** | **G1** | **G2** | **G3** | **G1** | **G2** | **G3** | **Total** |
| G1 |  | **44** |  |  | 4 | **40** | 44 |
| G2 | 8 | **51** |  | 7 | **43** | 9 | 59 |
| G3 |  | 6 | **16** | **21** |  | 1 | 22 |
| Total | 8 | 101 | 16 | 28 | 47 | 50 | 125 |

**Table 4** – Results from the ANOVA table of berry fruit data: P-values for main effects (0.000 means a P-value ≤ 0.0001), averages and standard errors for the product effect are reported. Sample codes are assigned according to the ingredient mixed with raspberry (A: apple, BO: blood orange, O: orange, P: pineapple and PG: pomegranate). Different letters indicate different HSD Tukey’s test homogenous groups.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **Sum Sq.** | | **d.f.** | | **Mean Sq.** | | **F** | | **Prob>F** |
| Products | 313.41 | | 4 | | 78.3514 | | 17.46 | | 0.000 |
| Consumers | 307.56 | | 71 | | 4.3318 | | 0.97 | | 0.559 |
| Error | 1274.19 | | 284 | | 4.4866 | |  | |  |
| **Product averages (standard error)** | | | | | | | | | |
| **A** | | **PG** | | **BO** | | **O** | | **P** | |
| 4.125a(0.244) | | 4.556ab(0.320) | | 5.472bc(0.218) | | 5.583c(0.252) | | 6.819d(0.190) | |

**Table 5 -** Cross tabulation of group assignment of automatic segmentation methods for raspberry data set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **HCA Complete** | | **K-means** | |  |
| **HCA Ward’s** | **G1** | **G2** | **G1** | **G2** | **Total** |
| **G1** | 20 | 18 | 24 | 14 | 38 |
| **G2** | 25 | 9 | 10 | 24 | 34 |
| **Total** | 45 | 27 | 34 | 38 | 72 |

**Figure captions**

**Figure 1** –Loadings (a) and scores (b) from PCA of the apple juice original data. The names in the score plot gives a direct indication on sample composition: the first letter indicating sugar level and the second letter indicating acid level (L: low, M: medium, and H: high).

**Figure 2** –Loadings with cluster indication (a) and scores (b) from PCA of the apple juice residuals. The names in the score plot gives a direct indication on sample composition: the first letter indicating sugar level and the second letter indicating acid level (L: low, M: medium, and H: high).

**Figure 3** – 95% confidence intervals of liking averages for the different products in the four groups from apple juice data set. The names in the score plot gives a direct indication on sample composition: the first letter indicating sugar level and the second letter indicating acid level (L: low, M: medium, and H: high).

**Figure 4** – PLS-DA plot for groups G1 and G3 of apple juice data set.

**Figure 5** – Comparison of observed explained variance with that obtained in the permutation test for apple juice (a) and raspberry data set (b). Note that 5 and 4 components are the maximum in the two cases and 100% is thus obvious. It is here only incorporated for comparison.

**Figure 6** – Visual comparison of complete linkage, K-means with 3 and 4 groups and visual grouping on residuals in PCA loading plots.

**Figure 7** –Loadings with cluster indication (a) and scores (b) from PCA of the raspberry residuals. Sample codes are assigned according to the ingredient mixed with raspberry (A: apple, BO: blood orange, O: orange, P: pineapple and PG: pomegranate).

**Figure 8** –Loadings (a) and scores (b) from PCA of the raspberry original data. Sample codes are assigned according to the ingredient mixed with raspberry (A: apple, BO: blood orange, O: orange, P: pineapple and PG: pomegranate). The arrow gives an indication of main preference direction.

**Figure 9** – 95% confidence intervals of liking average calculated for the different products in the three groups from raspberry data set. Sample codes are assigned according to the ingredient mixed with raspberry (A: apple, BO: blood orange, O: orange, P: pineapple and PG: pomegranate).

**Figure 10** – PLS-DA plot for the three groups of raspberry data set. Significant variables are circled.

**Figure 11 -** – Visual comparison of Ward’s linkage HCA and K-means with 2 groups in PCA loading plots.