Variable selection in PCA in sensory and consumer data

Frank Westad[1], Margrethe Hersleth[1], Per Lea[1] and Harald Martens[2]

[1]MATFORSK, Norwegian Food Research Institute, Osloveien 1, N-1430 Ås, Norway
[2]Sensory Science, The Royal Veterinary and Agricultural University, DK-1958 Frederiksberg, Denmark

In analysis of studies where sensory and consumer-related data are collected, it is important to extract the relevant information. One objective is to find significant sensory attributes, e.g. when interpreting loading plots from some factor analysis method. Whereas analysis of sensory data often gives interpretable factors and high percentage explained variance, consumer data tend to be less structured in terms of explained variance. For the consumers, many different groups of variables are usually present, such as preference, demography, eating habits, attitudes etc. These data often serve as a basis for segmentation of the consumers before preference mapping is employed. However, the segmentation should be validated, and removing non-relevant variables is usually more critical for these data than in sensory data.

A general approach for finding significant variables in multivariate models will be presented. The method is based on uncertainty estimates from cross-validation/jack-knifing [1], and the importance of model validation is emphasized. The approach has previously been employed in regression situations [2][3] [4], but is here focused on analysis of *one* set of data. Due to the rotational ambiguity in Principal Component Analysis (PCA), procrustes rotation is employed prior to estimating the uncertainties. Standard t-tests based on the loadings and their uncertainties give significance on each variable for each component. It will also be shown how this aids the data analyst in assessing the correct rank of the model.

One important question is how we can know if the estimates are close to the "truth". An example will be presented where a factorial design is the basis for eight different products. The objective was to compare ANOVA on the sensory attributes with variable selection in PCA. The results show that the same attributes are found to be significant, i.e. if some design variable is found to be significant on one attribute, then the PCA will give significance on one or more components for that attribute. This implies that the approach is relevant also for consumer data where the "truth" is not known. Other examples will be given of how to apply this method for segmentation and variable selection.

References

1. Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia, PA, (1982), ISBN 0-89971-179-7.
2. Martens, H. and Martens, M. (2000) Modified Jack-knife estimation of parameter uncertainty in bilinear modelling (PLSR). *Food Quality and Preference*, 6-15, **11**.
3. Martens, H. and Martens. M. (2001) Multivariate Analysis of Quality. An Introduction. J.Wiley & Sons, Ltd, Chichester UK.
4. Martens, H., Høy, M., Westad, F., Folkenberg, D. and Martens, M. (2001) Analysis of experiments by stabilised PLS Regression and jack-knifing. Chemometrics and Intelligent Laboratory Systems, 151-170, **58**.