Permutation testing for validating PCA

Isabella Endrizzi, Flavia Gasperi, Marit Rødbotten and Tormod Næs

Abstract Permutation test as possible alternative to the commonly used cross-validation of samples for validating PCA results is suggested. The approach is then illustrated using two data-sets from consumer studies of apple and raspberry juice. Our findings show that internal validation provided by the permutation test is particularly advantageous when the data are complex as they are in the second case reported here.

Keywords PCA, validation, permutation test, consumer test

1 Introduction

Internal preference mapping (Chang and Caroll, 1969; Caroll, 1972) is an important methodology based on principal component analysis (PCA), much used in consumer science for modelling, analysing and understanding consumer preferences. In this field we usually refer to a situation where a limited number of products are submitted to a

[1] Isabella Endrizzi, Department of Food Quality and Nutrition, Research and Innovation Centre, Fondazione Edmund Mach (FEM), Via E. Mach, 1, 38010 San Michele all'Adige, Italy ; email: isabella.endrizzi@fmach.it

Flavia Gasperi, Department of Food Quality and Nutrition, Research and Innovation Centre, Fondazione Edmund Mach (FEM), Via E. Mach, 1, 38010 San Michele all'Adige, Italy ; email: flavia.gasperi@fmach.it

Rødbotten Marit, 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; email: marit.rodbotten@gmail.com

Tormod Næs, 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; email: Tormod.Naes@nofima.no

consumer panel for the evaluation of the global liking. The sample set may be very small due to the fact that it is impossible to submit to the consumers a too high number of stimuli, according to sensory analysis basic rules. Therefore, the resulting data matrices are in the shape of a flat rectangle with a very low number of rows. The sample set is also often based on a designed experiment with possibly unique samples in it. Thus, if the practitioner is interested in validating the PCA on these data, the commonly used full cross-validation (CV) of samples it is probably not the best choice. Also because, the main purposes of validation in preference mapping is to determine the number of principal components that can safely be interpreted. Being the focus on interpretation of a given set of samples, the internal validation related to the actual samples at hand instead of some type of predictive performance of the model for other samples it may be more natural.

Here we propose as a possible alternative approach to use a permutation test for testing the significance of the principal components extracted from the consumer liking data. For illustrating the methodology we used two data sets from real consumer studies on apple and raspberry juice.

2 Permutation test

Permutation testing is a non-parametric tool which allows to evaluate 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 should produce a representative sampling distribution (Xiong, Blot, Meullenet and Dessirier, 2008). In the literature there are a number of methods available for permutation testing in PCA (Wakeling, Raats, and MacFie, 1991; Landgrebe, Wurst, and Welzl, 2002; Peres-Neto, Jackson, and Somers, 2003; Linting, van Os, and Meulman, 2011), but as far as we know, not for testing the significance of each additional component which is the focus here. The procedure used in this paper is described in the section below. Note that the interpretation is in terms of internal validity and different from regular CV for the samples.

2.1 Algorithm description

The measure used for testing significance of a new component is the explained variance relative to the sum 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. For a more detailed description see figure 1.

3 Data sets used

The first dataset has 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 (low, medium and high) and degree of acid (low and high). The prepared six juices were evaluated by a panel of 125 consumers who were asked to rate their degree of liking on a 7-point hedonic scale for each juice.

The second dataset has been used 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, Pirretti, Calò, and Gasperi, 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 asked to rate their appreciation on a 9-point scale.

4 Results

On both data sets a PCA was run, and in this the comparison of CV and permutation results are reported.

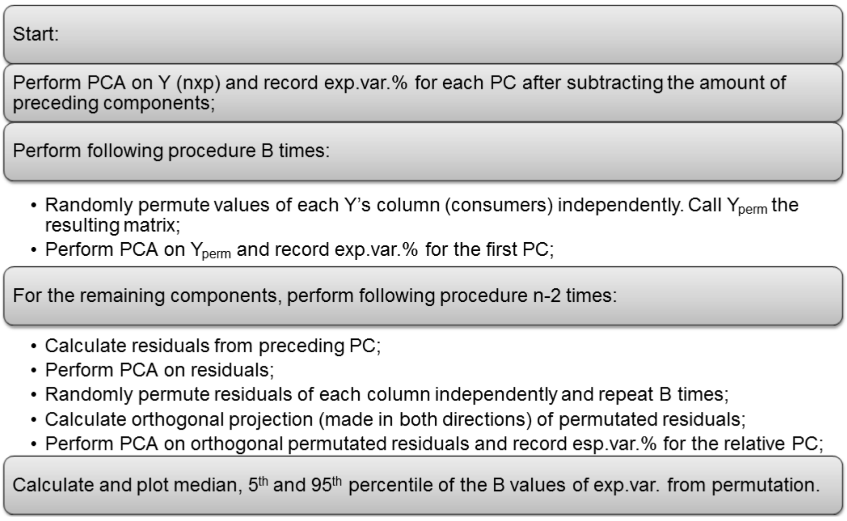
In table 1a, cumulated percentage of explained variances for each principal component using validation on samples for the apple juice data set are reported. The explained variance from CV of the samples clearly indicates two components as significant (only limited increase after), which corresponds well 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. On the same data set, the permutation test was run with B=300 in order to evaluate the significance of each component. In Figure 2a, 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 the explained variance relative to the sum of the variances for the remaining components as explained in section 2.1. 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 p1 = 0.010, p2 = 0.010, p3 = 0.218, p4 =0.802 confirming the conclusions given by the graph.

In table 1b, cumulated percentage of explained variances for each principal component using validation on samples 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 give no clear indication on the number of components to use. Using permutation test (Figure 2b), we have to conclude instead that the first three components are significant as confirmed by the corresponding P-values p1 = 0.010, p2 = 0.020, p3 = 0.010. Generalising, permutation test outperforms CV in data sets with a more complex experimental design with very different products.

Table 1: Cumulated percentage of explained variance in validation using samples 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.

PCs	*Val.Var%.(a)*	*Val.Var%.(b)*
PC_1	26.516	7.631
PC_2	35.626	6.530
PC_3	37.342	16.265
PC_4	37.089	- -

Figure 1: The permutation test algorithm for PCA



Start:

Perform PCA on Y (nxp) and record exp.var.% for each PC after subtracting the amount of preceding components;

Perform following procedure B times:

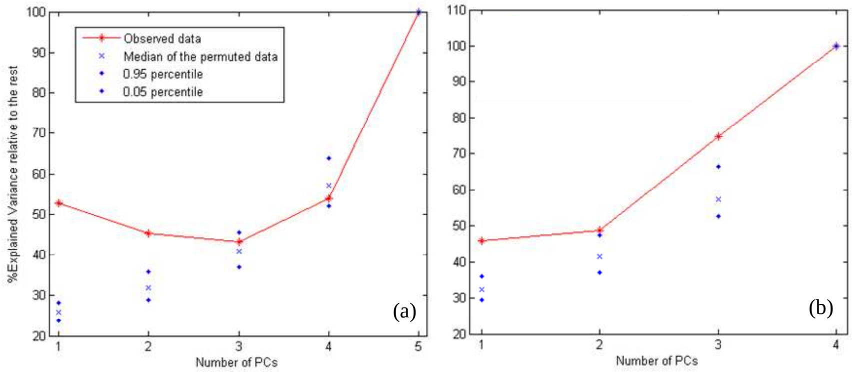
- Randomly permute values of each Y's column (consumers) independently. Call Y_{perm} the resulting matrix;
- Perform PCA on Y_{perm} and record exp.var.% for the first PC;

For the remaining components, perform following procedure n-2 times:

- Calculate residuals from preceding PC;
- Perform PCA on residuals;
- Randomly permute residuals of each column independently and repeat B times;
- Calculate orthogonal projection (made in both directions) of permutated residuals;
- Perform PCA on orthogonal permutated residuals and record esp.var.% for the relative PC;

Calculate and plot median, 5th and 95th percentile of the B values of exp.var. from permutation.

Figure 2: 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.



5 Conclusions

The authors recommend the use of the proposed permutation test which 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. The permutation test proposed here works well in any situation, but in comparison with the commonly used full cross-validation, a particular advantage of using it appears when the data are complex in terms of experimental design or strong variability as they are in the second case reported here.

References

1. Carroll, J.D.: Individual differences and multidimensional scaling. In: Shepard, R.N., Romney, A.K., Nerlove, S.B. (eds.) Multidimensional scaling: Theory and application in the behavioural sciences (Vol. 1), pp 105-155. Seminar Press, New York (1972)
2. Chang, J.J., Carroll, J.D.: How to use MDPREF, a computer program for multidimensional analysis of preference data (Tech. Rep.). Bell Telephone Laboratories, Murray Hill, NJ (1969)
3. Endrizzi, I., Pirretti, G., Calò, D.G., Gasperi, F.: A consumer study of fresh juices containing berry fruits. J. Sci. Food. Agr. {89}, 1227--1235 (2009)
4. Good, P.: Permutation tests: A practical guide to resempling methods for testing hypothesis (2nd ed.). Springer, New York (2000)
5. Landgrebe, J., Wurst, W., Welzl, G.: Permutation-validated principal components analysis of microarray data. Genome. Biol. {3}, 0019 (2002)
6. Linting, M., van Os, B.J., Meulman, J.J.: Statistical significance of the contribution of variables to the PCA solution: an alternative permutation strategy. Psychometrika. {76}, 440--460 (2011)
7. Peres-Neto, P.R., Jackson, D.A., Somers, K.M.: Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology. {84}, 2347--2363 (2003)

8. Rødbotten, M., Martinsen, B.K., Borge, G.I., Mortvedt, H.S., Knutsen, S.H., Lea, P., Næs, T.: A cross-cultural study of preference for apple juice with different sugar acid contents. Food. Qual. Prefer. {20}, 277--284 (2009)
9. Wakeling, I.N., Raats, M.M., MacFie, H.J.H.: A new significance test for consensus in generalized procrustes analysis. J. Sens. Stud. {7}, 91--96 (1992)
10. Xiong, R., Blot, K., Meullenet, J.F., Dessirier, J.M.: Permutation tests for Generalized Procrustes analysis. Food. Qual. Prefer. {19}, 146--155 (2008)