11th World Congress on Computational Mechanics (WCCM XI)
5th European Conference on Computational Mechanics (ECCM V)
6th European Conference on Computational Fluid Dynamics (ECFD VI)
July 20 - 25, 2014, Barcelona, Spain

THE UNCERTAINTY AND ROBUSTNESS OF THE PROCEDURES FOR THE DIMENSIONALITY REDUCTION

Jacek Pietraszek[1] and Ewa Skrzypczak-Pietraszek[2]

[1] Cracow University of Technology, Institute of Applied Informatics
Al. Jana Pawła II 37, 31-864 Kraków, Poland
pmpietra@mech.pk.edu.pl
[2] Jagiellonian University, Collegium Medicum, Chair and Department of Pharmaceutical Botany
ul. Medyczna 9, 30-688 Kraków, Poland
ewa.skrzypczak-pietraszek@uj.edu.pl

Key Words: *Principal Component Analysis, Cluster Analysis, Jackknife, Reduction of Dimensionality, Bootstrap, Melittis melissophyllum, Lamiaceae, Uncertainty, Robustness.*

The pharmacobotanical experimental studies usually provide data sets with relatively small number of records, but a rather large number of identified properties for each analysed sample of a plant or a tissue. The nature of such data lead them closer to the observational studies [1] rather than designed experiments [2]. It means that correlations and joint probability distributions including all observed properties are better tools than a regression implying a causal relationship.

The attempt to reduce the number of variables, required to describe the data unambiguously , is one of the first steps in the analysis of such sets. Such a procedure is called “the reduction of dimensionality” [3]. The typical approach is to consider the data set in a multi-dimensional space and to involve a geometric interpretation in such categories as clusters, hyper-planes, hyper-surfaces. Two popular methods are the cluster analysis (CA) [4] and the principal component analysis (PCA) [5].

CA is unsupervised technique that reveals the natural groupings existing between sample populations characterized by the values of a set of measures. Various types of the formal metrics are used but the most popular is a technique known as “the nearest neighbourhood” with a classic Euclidean distance. The typical form of the results is the dendrogram diagram. However, the interpretation of such diagram is highly subjective.

PCA is the multivariate method based on a covariance matrix. The method is one example from the set of the projection methods leading from the high-dimensional space containing data set onto a lower-dimensional subspace. PCA is eigenvalue-eigenvector problem. The eigenvectors are identified for the multi-dimensional “cloud” of data set samples and they are ordered according to descending eigenvalues being directly related to variances. The sorted plot of the eigenvalues (known as “scree plot”) is the main tool for selecting the most significance PCA factors explaining the main part of a total variance. Such selection is also highly subjective like in CA problem.

The CA and PCA techniques were used by authors for the pharmacobotanical analysis of phenolic acids [6] and flavonoids [7] in plants of *Melittis melissophyllum* L. (*Lamiaceae*) being old medicinal plant. The plant was considered as a source of raw material for biotechnological engineering. The obtained results, however fruitful and significant, revealed the necessity to identify the regions of confidence (ROC) for eigenvectors and confidence intervals (CI) for eigenvalues. Similar problem with uncertainty assessment was identified for CA.

The lack of information about probability distribution disabled the classic approach. The authors propose to analyse the uncertainty of the results by a bootstrap method [8] being a resampling kind of Monte Carlo approach. The method allows to obtain a full distribution of considered variable. The robustness of CA and PCA is analysed by a systematic deletion of a data subset and observing changes of the results.

REFERENCES

[1] J. Gentle, W.K. Härdle and Y. Mori, *Handbook of Computational Statistics*, Springer,2012.

[2] T.P. Ryan, *Modern Experimental Design*, John Wiley & Sons, 2007.

[3] A.J. Izenman, *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning*, Springer, 2008.

[4] B.S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis*, 5th Edition, John Wiley & Sons, 2011.

[5] I.T. Jolliffe, *Principal Component Analysis*, 2nd Edition, Springer, 2010.

[6] E. Skrzypczak-Pietraszek and J. Pietraszek, *Chemical profile and seasonal variation of phenolic acid content in bastard balm (Melittis melissophyllum L., Lamiaceae)*. J. Pharm. Biomed. Anal., Vol. **66**, pp. 154-161, 2012.

[7] E. Skrzypczak-Pietraszek and J. Pietraszek, *Seasonal Changes of Flavonoid Content in Melittis melissophyllum L. (Lamiaceae)*. Chem. & Biodiversity (accepted).

[8] J. Shao and D. Tu, *The Jackknife and Bootstrap*, Springer, 1995.