

KN3

REMARKS ON DATA CORRELATION

R. Todeschini, V. Consonni

*Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences,
University of Milano-Bicocca, Milano, Italy*

The correlation within a multivariate data set may influence the results of several statistical and chemometric methods. For instance, the parameters of regression models, the number of significant principal components in PCA and the composition of optimal informative subsets of objects in experimental design may change on the basis of the level of data correlation.

When using such methods it is thus essential to have a good estimate of how much data variability is related to systematic, and potentially useful, information rather than random noise and chance correlation. Indeed, random noise or chance correlation are always potentially present, making models both unstable and unreliable and giving undesired and sudden bias in data exploration.

In this presentation, some topics related to the concept of multivariate correlation, and specifically how to obtain information from correlated variables and the strategies to control and exploit this information in chemometrics [1-4], will be briefly introduced and discussed.

References

- [1] Todeschini, R., Data correlation, number of significant principal components and shape of molecules. The K correlation index. *Anal. Chim. Acta*, (1997) 348, 419-430.
- [2] Todeschini, R., Consonni, V. and Maiocchi, A., The K correlation index: theory development and its applications in chemometrics. *Chemometrics & Intell. Lab. Syst.*, (1999) 46, 13-29.
- [3] Todeschini, R., Consonni, V. and Pavan, M., A Distance Measure between Models: a Tool for Similarity/Diversity Analysis of Model Populations. *Chemometrics & Intell. Lab. Syst.*, (2004) 70, 55-61.
- [4] Todeschini, R., Consonni, V., Mauri, A. and Pavan, M., Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta*, (2004) 515, 199-208.