Editorial

Hans-Hermann Bock

Published online: 13 November 2009
© Springer-Verlag 2009

The current issue 3 of volume 3(2009) of the journal *Advances in Data Analysis and Classification (ADAC)* is a Special Issue that contains articles that were presented during the 11th Conference of the *International Federation of Classification Societies (IFCS)* that was held at the University of Dresden on March 13–18, 2009, with about 275 invited and contributed contributions and about 320 participants from over the world. These articles deal with structural and invariance properties in clustering, classification for data streams, robust dynamic graphics for mixture detection, problems related to correspondence analysis, and fraud detection with digit analysis.

François Brucker and *Alain Gély* introduce a new clustering structure with overlapping clusters, termed *parsimonious cluster system*, that generalizes phylogenetic trees and is related to weak hierarchies, binary dissimilarities, hypertrees, etc. Apart from various formal characterizations, it is shown that there are bijections between parsimonious clustering systems, a dissimilarity type termed *chordal quasi-ultrametrics*, and a known lattice structure (*dismantable lattices*). Suitable computation algorithms and graphical presentations are provided as well.

François Bavaud investigates ‘general clustering’ that involves weighted objects and fuzzy memberships and studies the (invariance) behaviour of classical and more recent clustering functionals when (more or less equivalent) groups or objects are aggregated. In addition to dispersion measures he considers convex clustering functionals and thermodynamic clustering and also detects a relationship between the likelihood analysis of a weighted mixture model and the determination of an optimum membership matrix by a modified EM algorithm.

Clustering of variables is the topic investigated by *Véronique Cariou*, *Stéphane Verdun*, *Emmanuelle Diaz*, *El Mostafa Qannari*, and *Evelyne Vigneau*. These authors

H.-H. Bock (✉)
Institute of Statistics, RWTH Aachen University, 52056 Aachen, Germany
e-mail: bock@stochastik.rwth-aachen.de

compare three statistical approaches for selecting an appropriate number of clusters of variables, using either a gap test, a weighted gap test, or a testing strategy based on a Clustering around Latent Variables (CLV) methodology. The method is applied and illustrated for the case of preference data.

When classifying objects into more or less known classes classical methods assume that for each object all data are available at the time of analysis. Recently, however, there are situations where data arrive sequentially in continuous streams and build up huge real-time data collections that cannot be stored simultaneously or re-analyzed at each time point, while objects must be classified in an online manner. This situation is considered in the paper by *Christoforos Anagnostopoulos, Dimitris K. Tasoulis, Niall M. Adams, and David J. Hand*: These authors show a method how to construct, in a logistic regression framework, online and temporally adaptive classifiers capable of handling the possibly drifting decision boundaries arising in streaming, dynamic or unstable environments.

The article by *Domenico Perrotta, Marco Riani, and Francesca Torti* considers the problem of identifying multiple masked outliers in a regression mixture context by using the forward search approach, i.e., by adapting the parametric model to an initial subset of the data, and then iteratively modifying (increasing) this subset in dependence on the observed residuals. Thereby, a large number of multiple views of hidden structures and a variety of plots emerge that should be evaluated simultaneously. This paper shows how these outputs can be usefully linked, presents a set of new robust graphical tools that support the statistician's analysis, and demonstrates its power on several practical regression problems.

Fraud detection is an important problem not only for banks, credit card enterprises, phone companies etc., but also for tax offices or accounting departments where subjects provide long lists of prices, income, taxable amounts, etc. that may be manipulated or faked and insofar may inadvertently violate some general statistical principles. The paper by *Stephan Dlugosz and Ulrich Müller-Funk* provides tools that use the (distribution of) values of the last, second last,... digits of the recorded numbers in a table of figures in order to find out if the list is suspect and likely to be provided by fraudsters. It is shown that, under various assumptions (different from Benford's law), the (asymptotic) distribution of last digits is uniform.

The article by *Jean Diatta* starts from the fact that many cluster collections obtained within the framework of cluster analysis are Moore families, i.e., the intersection of two clusters belongs to the family again (or is empty). Moore families are closely related to *implicational systems* that are well-known as attribute implications in Data Mining and market basket research, or as functional dependencies in relational databases. A minimal set of implications of an implicational system (a Guigues-Duquenne basis) is characterized in terms of particular quasi-closed sets, the 'critical sets'. The paper provides a structural analysis of this field, e.g., a simple characterization of critical sets, and investigates the case of a weakly hierarchical cluster structure. Moreover, an efficient algorithm for finding critical sets is given.

Bernard Fichet starts with the Principle of Distributional Equivalence (P.D.E.) in correspondence analysis where the dissimilarity among columns of the underlying contingency table is measured by the χ^2 metric. He asks if the P.D.E. property may be maintained when measuring dissimilarity by some L_p -type metric between the

columns that involves a certain weight function g . His paper specifies conditions for g such that P.D.E. is maintained, and for the existence of such a g . Finally he presents some numerical examples.

The last paper of this issue, written by *Yoshio Takane*, is situated in the framework of Nonsymmetric Correspondence Analysis (NCA) where we look for predictive relationships between rows and columns of a contingency table, and considers the case when some auxiliary information on the predictor categories (rows, say) is available. This information may be integrated in the form of linear constraints (constrained NCA), but this may leave unaccounted for a part of the predictive relationship. The paper presents a method for analyzing the residual part along with the part accounted for by the constraints, both being tested for significance. The paper shows that this procedure parallels the distinction between tests of ignoring, and tests of eliminating, and also establishes some relationship to Simpson's paradox.