Editorial

The final issue of the *Journal of Classification* for 2016 contains seven articles that extend advances in the field of classification. The first paper is a review paper of Model-Based clustering by Paul McNicholas. McNicholas, who is quickly becoming one of the world's leading experts on model-based clustering (see McNicholas, 2016, for his monograph level treatment of the subject), provides a real treat in this overview of the field. Model-based clustering has been a topic of interest since the *Journal of Classification's* inception (Basford and McLachlan, 1985); recently, along with several other papers utilizing mixture models (Evans, Love and Thurston, 2015; Marbac, Biernacki and Vandewalle, 2015, for instance), he has contributed two developments in the area to the *Journal of Classification* (Andrews and McNicholas, 2014; Vrbik and McNicholas, 2015). The review begins with seminal work in the field, especially that of Wolfe (1963; 1965), proceeds through the ground-breaking eigen-decomposition of Banfield and Raftery (1993), and ends with modern developments in non-Gaussian mixtures, asymmetric mixtures, dimension reduction, and the like. The paper contains numerous references and can easily serve as a keystone for graduate students or researchers in the field to quickly access many of the most important and relevant domains of model-based clustering.

McNicholas' review is followed by a paper developing a regression mixture approach for clustering functional data by Faciel Chamroukhi. This approach to functional data analysis combines both the traditional "fuzzy" mixture modeling where posterior probabilities are estimated and the classification mixture model where observations are partitioned in a "hard" fashion (e.g., they are either in a group or are not in a group and they can only belong to one group). Dynamic programming is used to optimally segment the functional curves, a task it is especially suited for given the ordered nature of the data (see Steinley and Hubert, 2008). Further, Chamroukhi shows that his method is a generalization of standard K-means clustering algorithm for segmenting curves and performs better in many situations evaluated within the simulation and analysis of real data sets. While functional data analysis has seen some advances in the *Journal of Classification*, most recently Alonso et al. (2014), the advancing nature of modern data—both analysis and collection—predicts that this topic will become increasingly popular.

The third paper, by Semhar Michael and Volodymyr Melnykov, provides a nice follow-up to the first two articles in that it is the advancement of model-based clustering of regression time series using both the Kalman filter to obtain exact maximum likelihood estimates and an alternative, computationally efficient approach for larger time series.

These advances follow some prior work by Melnykov (2012) in efficiently estimating regression time series. Finally, a convincing example is provided where the models are applied to dendrochronology (e.g., tree ring) data.

Michael Greenacre and Patrick Groenen provide a sharp departure from mixture models with their presentation on weighted Euclidean biplots. As a type of multivariate representation, biplots have seen frequent exposure in the *Journal of Classification* (see Underhill, 1990, for an early example). Greenacre and Groenen develop a weighted least squares approximation that has, among other advantages, the property that any non-Euclidean proximity can be approximated by a Euclidean counterpart, allowing for the straight forward interpretation of the represented distances and the contributions of the original variables. This is all achieved through the machinations of the singular-value decomposition, leading to, as one would hope, the proposed method being a generalization of both principal components analysis and correspondence analysis.

As noted in the Editorial of the last issue of the *Journal of Classification*, Cognitive Diagnosis Models are becoming increasingly popular, and Chen-Yu Hong, Yu-Wei Chang, and Rung-Ching Tsai continue the trend in this issue by introducing order restrictions to the DINA model (and, from the last issue, we know that the DINA model and the DINO model share a duality, so the results presented should also apply to the DINO model; see Kohn and Chiu, 2016). The current paper solves a problem that I have long been vexed by—namely, the fact that in the traditional DINA model an examinee who had mastered more required attributes could have a lower probability of getting a correct answer than an examinee who had mastered fewer required attributes. Furthermore, an added benefit of the provided estimation method is violation of this restriction can indicate misspecification in the so-called Q-matrix. As the field of CDM moves forward, it is nice (and welcome) to see increased emphasis placed on model diagnostics and internal consistency.

The penultimate paper in this issue by Sancar Adali and Carey Priebe focuses on the trade-off between the preservation of within-condition dissimilarities and between-condition similarities when embedding observations from two different data sources. Specifically, they have developed a new multidimensional scaling approach that handles this situation, allowing the end-user to determine the trade-off they wish to focus on (e.g., equal weighting, optimal weighting, or some other weighting scheme). The current approach is similar to individual differences scaling (Carroll and Chang, 1970); however, this approach does not require the estimation of a “common” space. Herein, the goal is to “match” observations across data sources, which is basically a form of pattern recognition. Personally, I think an interesting extension would be the combination of the methods presented herein with the notions laid

forth in van Buuren and Heiser (1989) to find clusters of individuals in reduced space that are matched across data sources, allowing the relaxation of matching individual observations and instead to match groups of observations.

The final article is one by Matthijs Warrens and Bunga Pratiwi devoted to defining Kappa coefficients of agreement for circular data. The authors first show that traditional Kappa coefficients designed for linear data are deficient for measuring the agreement between two circular classifications and then proceed to develop a class of circular Kappas, providing a set of proofs that highlight many of their properties. Further, they show that these circular Kappas tend to produce higher values, so if one were to use the traditional cutoffs—which should likely be advised against, as most cutoffs are arbitrary—then care should be taken to adjust one's preconceptions accordingly.

Douglas Steinley

References

- ALONSO, A. M., CASADO, D., LOPEZ-PINTADO, S., and ROMO, J. (2014), “Robust Functional Supervised Classification for Time Series”, *Journal of Classification*, 31(3), 325–350.
- ANDREWS, J. L., and MCNICHOLAS, P. D. (2014), “Variable Selection for Clustering and Classification”, *Journal of Classification*, 31(2), 136–153.
- BANFIELD, J. D., and RAFTERY, A. E. (1993), “Model-Based Gaussian and Non-Gaussian Clustering”, *Biometrics*, 49(3), 803–821.
- BASFORD, K. E., and MCLACHLAN, G. J. (1985), “The Mixture Method of Clustering Applied to Three-Way Data”, *Journal of Classification*, 2(1), 109–126.
- CARROLL, J. D., and CHANG, J.-J. (1970), “Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of ‘Eckart-Young’ Decomposition”, *Psychometrika*, 35(3), 283–319.
- EVANS, K., LOVE, T., and THURSTON, S. W. (2015), “Outlier Identification in Model-Based Cluster Analysis”, *Journal of Classification*, 32(1), 63–84.
- KOHN, H.-F., and CHIU, C.-Y. (2016), “A Proof of the Duality of the DINA Model and the DINO Model”, *Journal of Classification*, 33(2), 171–184.
- MARBAC, M., BIERNACKI, C., and VANDEWALLE, V. (2015), “Model-Based Clustering for Conditionally Correlated Categorical Data”, *Journal of Classification*, 32(2), 145–175.
- MCNICHOLAS, P. D. (2016), *Mixture Model-Based Classification*, Boca Raton, FL: Chapman & Hall/CRC Press.
- MELNYKOV, V. (2012), “Efficient Estimation of Model-Based Clustering of Gaussian Regression Time Series”, *Statistical Analysis and Data Mining*, 5, 95–99.
- STEINLEY, D., and HUBERT, L. (2008), “Order Constrained Solutions in K-means Clustering: Even Better Than Being Globally Optimal”, *Psychometrika*, 73(4), 647–664.
- UNDERHILL, L. G. (1990). “The Coefficient of Variation Biplot”, *Journal of Classification*, 7(2), 241–256.
- VAN BUUREN, S., and HEISER, W. (1989), “Clustering N Objects into K Groups Under Optimal Scaling of Variables”, *Psychometrika*, 54(4), 699–706.

- VRBIK, I., and MCNICHOLAS, P. D. (2015), "Fractionally-Supervised Classification", *Journal of Classification*, 32(3), 359–381.
- WOLFE, J. H. (1963), "Object Cluster Analysis of Social Areas", Master's thesis, University of California, Berkeley.
- WOLFE, J. H. (1965), "A Computer Program for the Maximum Likelihood Analysis of Types", Technical Bulletin 65-15, U. S. Naval Personnel Research Activity.