Editorial

The first paper in this issue of the *Journal of Classification* focuses on the growingly popular cognitive diagnosis model (CDM) found in educational assessment (in fact, five of the six articles in the special issue of this journal, Volume 30(2), were dedicated to CDMs). While short in length, I do not believe the message of this article could be any more important. Specifically, Hans-Friedrich Köhn and Chia-Yi Chiu provide a proof that shows that two of the most popular CDMs (Henson, Templin, and Willse 2009, for a review of CDMs), the DINO and the DINA model, give equivalent results because, up to a linear transformation, they are identical models. As the authors note in their conclusion, this consolidation of methods greatly benefits the field as it obviates the need for numerical studies that focus on both methods (either the same article containing both methods, or somewhat worse, separate articles for each method); similarly, the result holds for proofs as well—what holds for one model also holds for the other. I see this result as a substantial step in the on-going maturation of CDMs.

The second paper, by Donniell Fishkind, Cencheng Shen, Youngser Park, and Cary Priebe, highlights the incommensurability phenomenon in principal components analysis (PCA). Succinctly, the authors demonstrate that the data reduction, via PCA, on two separate data sets that are believed to be random samples from the same population can lead to wildly different results in terms of the projected data. Furthermore, and seemingly counterintuitively, this can even be the case when the two original data sets are highly correlated. The authors note that this can cause problems in several situations and illustrate issues that can arise in distributed settings (e.g., merging highly correlated data sets after projection); however, as noted in the introduction, the notion of incommensurability extends to other data reduction techniques, such as multidimensional scaling. It is completely feasible that the same problem exists in the context of factor analysis and the inability to replicate findings in the behavioral and health sciences (Simmons, Nelson, and Simonsohn 2011) using these models (and by extension, structural equation models). Further, this is potentially exacerbated in these settings as the correlation between variables within a data set, and likely between data sets measuring the same variables, is much lower than the idealized setting of highly correlated data sets. If the problem of incommensurability is readily present in the latter, it should be expected that it is exhibited to a larger degree in the former.

The third paper by Renato Cordeiro de Amorim addresses variable (e.g., feature) weighting in the context of K -means clustering and follows

the same author's work on feature weighting in Ward's hierarchical clustering (de Amorim 2015). Whereas the latter paper developed a new feature weighting method specifically for Ward's clustering, the current manuscript focuses on a comparison of existing variable weighting methods for K-means clustering. The simulation compares nine different methods, from classic methods such as SYNCLUS (DeSarbo, Carroll, Clark and Green 1984) and more recent methods such as that combine choosing the metric space and feature weighting simultaneously (de Amorim and Mirkin 2012). As with any simulation, some methods were not included in the comparison, such as the more promising methods found in Steinley and Brusco (2008a) that rely on notions set forth by Brusco and Cradit (2001) and Steinley and Brusco (2008b). Given the conclusion of recommendations by de Amorim regarding feature selection, I believe this area of research will prove to be fertile for some time as developing effective algorithms that meet multiple of those requirements (let alone all of them) will likely prove quite challenging.

The next paper by Michail Tsagris, Simon Preston, and Andrew Wood provides a paper that improves classification of compositional data (e.g., vectors whose non-negative elements sum to unity). The authors develop an intermediary transformation for compositional data that lies between what they term Euclidean Distance Analysis and Logratio Analysis and compare the proposed method and existing method in terms of classification performance in regularized discriminant analysis and k nearest neighbor classification. In the article, the authors advocate that rather than adhering to any arbitrary, theoretical mathematical properties for a particular transformation that the "proof is in the pudding" and all that matters is the final classification accuracy. Thus, for a given data set, the optimal transformation (in terms of minimizing classification error) can be chosen, highlighting the gains that can be obtained from "letting the data speak."

The fifth paper has Carlo Grillenzoni enhancing the blurring mean-shift algorithm for finding clusters in either spatial or image data. The algorithm is based on the kernel density function (Silverman 1986) and is iterative in nature. Like many more complex clustering techniques such as simulated annealing, genetic algorithms, and the like, there are many user-defined parameters that must be provided prior to the implementation of the algorithm. For kernel density estimation, one such parameter is the bandwidth for estimating the density. Specifically, if the bandwidth is too large, only one cluster will be found; conversely, if the bandwidth is too small, too many clusters will be uncovered. Grillenzoni provides an automatic, data driven criteria to select the bandwidth and lead to (hopefully) detecting the correct number of clusters.

The penultimate article is a theoretical treatment by Elena Rubei on weighted graphs and trees. Mathematical development with regard to tree

structures has a long history in the *Journal of Classification*, dating back to its very first issue (see Brown and Day 1984). Rubel's paper adds to the legacy by developing the identification of 2-weights for weighted graphs for fixed ranges, allowing for the capturing of a variability within a process that generates a weighted graph.

The final paper in this issue is from Roberta Siciliano, Antonio D'Ambrosio, Massimo Aria, and Sonia Amodio and focuses on analyzing the history of visiting web sites through different distance-based visualization methods of multidimensional scaling and correspondence analysis, both of which have a rich history in the *Journal of Classification* (see Choulakian and de Tibeiro 2013; Ma, Cardinal-Stakenas, Park, Trosset, and Priebe 2010, for more recent offerings). The authors are able to represent the structure of the web-history by creating a symmetric similarity matrix from the web-preference matrix and its component matrices (which can be thought of as similar to an affiliation matrix in network analysis), an approach that is not unlike those used in the blockmodeling literature for two-mode binary networks (see Brusco and Steinley 2007). Siciliano et al. close with a series of examples where this analytic approach could be put to fruitful use.

Douglas Steinley

References

- BROWN, E.K., and DAY, W.H.E. (1984), "A Computationally Efficient Approximation to the Nearest Neighbor Interchange Metric", *Journal of Classification*, 1(1), 93–124.
- BRUSCO, M.J., and CRADIT, J.D. (2001), "A Variable Selection Heuristic for K-means Clustering", *Psychometrika*, 66(2), 249–270.
- BRUSCO, M.J., and STEINLEY, D. (2007), "A Variable Neighborhood Search Method for Generalized Blockmodeling of Two-Mode Binary Matrices", *Journal of Mathematical Psychology*, 51(5), 325–338.
- CHOULAKIAN, V., and DE TIBEIRO, J. (2013), "Graph Partitioning by Correspondence Analysis and Taxicab Correspondence Analysis", *Journal of Classification*, 30(3), 397–427.
- DE AMORIM, R.C. (2015), "Feature Relevance in Ward's Hierarchical Clustering Using the Lp Norm", *Journal of Classification*, 32(1), 46–62.
- DE SARBO, W.S., CARROLL, J.D., CLARK, L.A., and GREEN, P.E. (1984), "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables", *Psychometrika*, 49(1), 57–78.
- MA, Z., CARDINAL-STAKENAS, A., PARK, Y., TROSSET, M.W., and PRIEBE, C. (2010), "Dimensionality Reduction on the Cartesian Product of Embeddings of Multiple Dissimilarity Matrices", *Journal of Classification*, 27 (3), 307–321.
- SILVERMAN, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.
- SIMMONS, J.P., NELSON, L.D., and SIMONSON, U. (2011), "False Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", *Psychological Science*, 22, 1359–1366.

- STEINLEY, D., and BRUSCO, M.J. (2008a), "Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures", *Psychometrika*, 73(1), 125–144.
- STEINLEY, D., and BRUSCO, M.J. (2008b), "A New Variable Weighting and Selection Procedure for K-means Cluster Analysis", *Multivariate Behavioral Research*, 43(1), 77–108.