Reviewer's report on doctoral thesis:

J. Vávra: Model-based Clustering of Multivariate Longitudinal Data of a Mixed Type

The main achievement of the dissertation is the description, implementation, and application of two approaches to the so-called *model-based clustering* (MBC) of *mixed-type* multivariate longitudinal observations.

Compared to existing literature concerning clustering of longitudinal data *[McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of longitudinal data. Canadian Journal of Statistics, 38(1), 153-168, Teuling, N. D., Pauws, S., & Heuvel, E. V. D. (2021). Clustering of longitudinal data: A tutorial on a variety of approaches. arXiv preprint arXiv:2111.05469, among others]* and its extension to multivariate longitudinal data *[Maruotti, A., & Punzo, A. (2017). Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. Computational Statistics & Data Analysis, 113, 475-496, Zhou, J., Zhang, Y., & Tu, W. (2022). clusterMLD: An Efficient Hierarchical Clustering Method for Multivariate Longitudinal Data. Journal of Computational and Graphical Statistics, (just-accepted), 1-36, Lim, Y., Cheung, Y. K., & Oh, H. S. (2020). A generalization of functional clustering for discrete multivariate longitudinal data. Statistical Methods in Medical Research, 29(11), 3205-3217]*, this dissertation allows to include both categorical and numerical observations — possibly with the exception of the very recent paper *[Tan, Z., Shen, C., Subbarao, P., Lou, W., & Lu, Z. (2022). A Joint Modeling Approach for Clustering Mixed-Type Multivariate Longitudinal Data: Application to the CHILD Cohort Study. arXiv preprint arXiv:2210.08385]* that also seems to be applicable for longitudinal data of a mixed type.

Contents

Chapter 1 introduces the longitudinal setup with mixed-type multivariate observations, followed by a description of two real-life data sets (PBC and EU-SILC). Models for correlated data (i.e., mixed-effects models) are described in Chapter 2, including two possible approaches for modelling categorial observations; the presentation of models uses conditional distributions. Classical model-based clustering is introduced in Sections 3.1 and 3.2. Section 3.3 reviews existing results for model-based clustering of longitudinal data. The author also presents an application of MBC to the PBC data set (using both the threshold and GLMM model for categorical variables), although the details concerning its implementation are postponed to Chapters 5–7.

Bayesian approach to parameter estimation is proposed in Chapter 4, including discussion of prior distributions. This chapter starts with listing reasons for not using maximum likelihood approach but, in my opinion, some of these problems (not considering MBC) have already been addressed, e.g., in the Software for Analysis of Binary and Recurrent Events (SABRE, `http://sabre.lancs.ac.uk/`), see also *[Berridge, Crouchley, & Grose (2011).* Multivariate generalized linear mixed models using R. *Boca Raton: CRC Press.]*

After suggesting tractable prior distributions, the author recalls Bayesian data augmentation and explains the principles of the MCMC approach. The structure of the two proposed MBC models is plotted in Figures 4.3 and 4.4.

Chapters 5 and 6 contain all technical details (concerning the posterior distribution) necessary for the implementation of the MCMC algorithm both for the threshold and GLMM approach to categorical variables. In both chapters, a simulation study investigates the performance of the Bayesian estimators but, unfortunately, these two simulation studies are using different designs that do not allow a direct comparison of the two approaches.

Further details concerning the software implementation in R (and C) are given in Chapter 7, including the analysis of PBC data set as an example. Finally, Chapter 8 contains an application of the GLMM approach to EU-SILC data set – the authors identifies 6 or 11 groups that mostly seem to be very similar from the point of view of income (Figure 8.1) but exhibit different time development (Figure 8.2).

Comments

The dissertation is written very clearly and it contains only a very small number of typing errors. In my opinion, the main contribution is the development and software implementation of MCMC algorithms allowing the implementation of MBC within the framework of *mixed-type* multivarite longitudinal observations.

Comments and questions concerning the methodology:

1. The proposed model-based techniques assume validity of the underlying models. Which techniques can be used to detect lack of fit? What happens if the underlying models are misspecified?

2. The strict distinction between the "threshold" and the "GLMM" model seems to be artifical. Would it be possible to combine these approaches in a single model, i.e., to model some categorical variables with the threshold model and some categorical variables with the GLMM model? Are there any guidelines for choosing one of these models?

3. What are the main differences compared to *[Tan, Z., Shen, C., Subbarao, P., Lou, W., & Lu, Z. (2022). A Joint Modeling Approach for Clustering Mixed-Type Multivariate Longitudinal Data: Application to the CHILD Cohort Study. arXiv preprint arXiv:2210.08385]*? Please, comment on the R package `BCClong` (Bayesian consensus clustering for multivariate mixed-type longitudinal data).

Some additional questions and comments:

p. 25: Typing error in formula (2.2).

p. 33 Is it possible to use uniform distribution (instead of normal) in the threshold model? Wouldn't it be easier to interpret?

p. 54, p.134: Covariance matrices of random effects are always general positive-definite matrices? Is it possible to assume (test?) that these covariance matrix have a simpler structure? Can this be implemented in the Bayesian approach?

Chapters 5–6 I miss a direct comparison of the "threshold" and "GLMM" approach in a single simulation study. Which of these approaches is better?

p. 134 Instead of extending the results to an even larger number of outcomes, it might be easier to reduce the dimensionality of the original data set, e.g., by applying principal components.

In my opinion, the MBC clustering of mixed-type longitudinal data is a challenging estimation problem that was addressed (and solved) very well in this dissertation. I fully agree that 'there are many possible directions for the future research to further enhance the methodology developed in this thesis' (p. 134) — certainly, the precise direction of the future research will strongly depend on the real-life problem under investigation — see, for example *[Lu, Z., & Lou, W. (2022). Bayesian consensus clustering for multivariate longitudinal data. Statistics in Medicine, 41(1), 108-127]* or *[Park, J., & Ahn, J. (2017). Clustering multivariate functional data with phase variation. Biometrics, 73(1), 324-333.]* Generalizations to multivariate functional observations may be another very interesting direction for the future research, see, for example *[Jacques, J., & Preda, C. (2014). Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis, 71, 92-106, Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. Computational Statistics, 35(3), 1101-1131.]*

Summary

The dissertation extends the MBC technique to mixed-type longitudinal data, including also the entire mathematical background necessary for implementing the MCMC algorithm in the statistical software R. The resulting methodology is directly applicable in practice; I appreciate also the software implementation and the interesting applications in economics (EU-SILC) and medicine (PBC). Altogether, this dissertation clearly demonstrates the author's capability to independent creative work and, therefore, I recommend to award the scientific degree Ph.D. to Jan Vávra.

Doc. RNDr. Zdeněk Hlávka, Ph.D.