WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Vienna, 21.12.2022

Evaluation of the Doctoral thesis of Jan Vávra on

„Model-based Clustering of Multivariate Longitudinal data of a Mixed Type"

The past decades have seen a tremendous increase in the availability of longitudinal data in many areas of applied research, such as marketing, economics, and finance, both in the life sciences, demography and environmental research. For a long time, the observed outcomes have been treated as continuous data and were modelled as arising from Gaussian distributions, even if the data were small counts or categorical variables. More recently, appropriate modelling of these kind of mixed-type data has become an important area of research in the statistical sciences.

Appropriate statistical modelling of multivariate longitudinal data is confronted with a number of challenges. First, the observations for a single unit are correlated and valid modelling has to consider this dependence. Second, the outcomes often are of mixed type and joint modelling of potentially highly correlated outcomes of different type is far from obvious. Third, the data might arise from an inhomogeneous population where the effects of the observed covariates differ across unknown subpopulations. Fourth, relevant covariates might be missing and random effects have to be introduced to account for unobserved heterogeneity. And last, but not least, statistical inference for these models is very challenging.

The candidate Jan Vávra has submitted an excellent thesis that addresses these concerns. The thesis suggests very innovative and up-to-date solutions and provides an important step forward in this highly relevant research area.

Longitudinal data are introduced and two real data examples are introduced in Chapter 1.

In Chapter 2, a unifying approach for joint modelling of highly dependent mixed type data is discussed which is based on the thresholding concept. This elegant concept relates the observed discrete or categorical outcomes to continuous latent variables. Dependence among the observed outcomes in introduced by joint modeling of the latent variables. In addition, generalized linear mixed-models are applied to allow for random-effects to account for missing covariates.

In Chapter 3, model-based clustering based on mixture models is employed to address the issue of unobserved heterogeneity. As a major advantage of this approach, the data are divided in subpopulations, which are determined endogenously from the data. In each subpopulation, one of the models introduced in Chapter 2 is applied as segment-specific model.

DEPARTMENT OF FINANCE,
ACCOUNTING & STATISTICS

Institute for Statistics and Mathematics

UNIV.PROF. DIPL.-ING. DR. SYLVIA FRÜHWIRTH-SCHNATTER
T +43-1-313 36-5581, F +43-1-313 36-90-5581
Welthandelsplatz 1, Building D4, 1020 Vienna, Austria
sylvia.fruehwirth-schnatter@wu.ac.at, www.wu.ac.at/statmath

In Chapter 4, modern Bayesian inference is applied to perform statistical inference in practice. This approach has the advantage that all unknown parameters, including all latent variables can be estimated jointly. Two challenges come with the Bayesian approach. First, this technique requires the choice of appropriate prior distributions to ensure a well-defined posterior distribution, which is the bases for further inference. Prior choices are particularly relevant for mixture models and are discussed in details in this chapter.

Second, the application of Bayesian inference requires the design of appropriate Markov chain Monte Carlo (MCMC) methods. The development of efficient MCMC samplers is very challenging for the complex models considered in this thesis. Only carefully designed sampler will be able to fully explore the posterior distribution. New sampling techniques are developed in the thesis and details are outlined for two models in Chapter 5 and 6. In Chapter 5, Gibbs sampling is implemented for the threshold concept model. The Gibbs sampler has the advantage to work without any tuning, however it might be prone to slow mixing. In Chapter 6, Metropolis within Gibbs (MH) sampling is implemented for GLMM based models. MH sampling requires the choice of appropriate proposal densities. This again is a challenging problem in high-dimensional parameter spaces and Taylor expansions are used in the thesis for this purpose.

Chapter 7 discusses in detail the software implementation for the models and methods introduced in the thesis. New solutions for several important practical aspects are suggested, such as deriving the classification probabilities and dealing with label switching.

Finally, the thesis concludes in Chapter 8 with a detailed analysis of data from the EU-SILC database on living conditions of Czech households.

Overall, Jan Vávra is to be congratulated on submitting a thesis of excellent scientific quality and I highly recommend the Faculty of Mathematics and Physics at Charles University, Prague, to award the candidate the desired degree.

Yours sincerely,

Univ.Prof.Dr.Sylvia Frühwirth-Schnatter