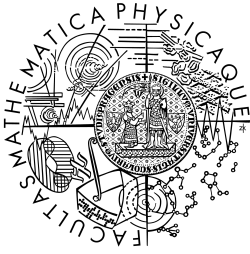Supervisor's report on the Ph.D. study and dissertation thesis by

Jan Vávra

Model-based Clustering of Multivariate Longitudinal Data of a Mixed Type

Jan Vávra has been studying Ph.D. program *Probability and Mathematical Statistics, Econometrics and Financial Mathematics* since autumn 2018. Before that, he had successfully completed both bachelor and master studies at MFF UK in a related study branch of *Probability, Mathematical Statistics and Econometrics* where he completed the diploma thesis also under my supervision. Already during this pre-doctoral period, Jan has shown pretty responsible attitude towards his study duties and developed very sound theoretical background in statistics which allowed him to start working on his doctoral research quite efficiently from the very beginning. During his doctoral period, it appeared that Jan is strong not only in theory but also in computational aspects of statistical methods which is also clearly demonstrated in his dissertation thesis.

The topic of the thesis is subject of an active research in several research groups around the world. It combines three research areas: (multivariate) longitudinal data that became largely available in many application areas in past decades, methods for categorical data and finally, methods of classification/clustering. As it is also illustrated in the introductory parts of the thesis, the portfolio of proper statistical methods is still relatively sparse when someone needs to analyze multivariate longitudinal data, i.e. longitudinal data where more than one outcome is being observed at each measurement occasion and on top of that, some/all outcomes are of a categorical nature. The situation becomes even more complex (and available methods even more rare) when such data are to be used for (unsupervised) classification/clustering. The thesis starts with four preparatory chapters and then provides (in key Chapters 5 and 6) two classes of methods to allow for that. The Jan's research aimed to develop proper statistical methods not only on a theoretical basis but also make them available to practitioners. Due to complexity of studied models, Jan opted for Bayesian methods which on one hand are capable to deal with even relatively complex models. Nevertheless, on the other hand, they require development and also implementation of computational algorithms towards Markov chain Monte Carlo estimation of such models. This part of Jan's doctoral research is summarized in Chapter 7 of the thesis and most importantly, it is also available on a `GitHub` as a (not small) set of computer programs coded in C/C++ (core computational routines) and R (user-friendly interface). Practical applicability of developed methods is illustrated throughout the thesis in several simulation studies and analyses of real datasets, the most importantly on the analysis of the Czech subset of the European EU-SILC database in Chapter 8.

Arnošt Komárek, Ph.D.
Associate Professor of Statistics
E-mail: komarek@karlin.mff.cuni.cz
Tel.: (+420) 951 553 282

Dept. of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University
Sokolovská 49/83, 186 75 Praha 8 – Karlín
Czech Republic

The key parts of the thesis are based on two research papers. The first one has appeared in *Advances in Data Analysis and Classification* (Q_2 by both IF and AIS in *Statistics and Probability* category). The second paper has successfully passed the first review round in *Statistics and Computing* (Q_2 by IF and Q_1 by AIS) with rather positive comments of both referees and I have no doubts that the revision will be accepted for publication soon after the defense. On top of those two papers, Jan published shorter reports in several conference proceedings (some of them indexed also on the *Web of Science*). He participated to several international as well as national scientific meetings. His work was awarded by the 1st place at the competition of Ph.D. students at the *Mathematical Methods in Economics* conference in 2021 (organized by Czech Society for Operational Research) and by the shared 2nd place at conference *ROBUST* 2022 (organized by Czech Statistical Society).

In summary, both the thesis and the course of Jan's doctoral studies proves that he is able to conduct a solid and independent research in statistics and related fields. That is, I have no doubts that the thesis fulfills all the requirements for the doctoral dissertation at the Ph.D. program *Probability and Mathematical Statistics, Econometrics and Financial Mathematics* at MFF UK and I **strongly recommend** it for the defense.

doc. RNDr. Arnošt Komárek, Ph.D.
KPMS MFF UK
22 December 2022

Arnošt Komárek, Ph.D.
Associate Professor of Statistics
E-mail: komarek@karlin.mff.cuni.cz
Tel.: (+420) 951 553 282

Dept. of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University
Sokolovská 49/83, 186 75 Praha 8 – Karlín
Czech Republic