Dynamic clustering

Ernst, Anja Franziska

DOI:
10.33612/diss.196176258

Publication date:
2022

Chapter 6

Conclusion and discussion

The aim of this dissertation was to develop novel dynamic clustering procedures to describe and uncover between-individual differences in intensive longitudinal data. Dynamic clustering procedures seek to identifying subgroups of individuals with similar dynamics, a priori these subgroups are unknown. By identifying homogeneous subgroups several individuals can be incorporated into the same longitudinal model, while still accounting for between-individual differences in the underlying dynamics. This final chapter summarises the developments of the dynamic clustering procedures that were presented in this dissertation, compares their estimation procedures, discusses their empirical applicability, and specifies avenues for future research.

6.1 Model specification and estimation

In this chapter we refer to a dynamic clustering procedure as a dynamic clustering model when it is a holistic model that integrates the two characteristic elements, summarising individuals' dynamics and clustering, into a single statistical model. A dynamic clustering procedure that does not integrate these two elements into a single statistical model is referred to as a dynamic clustering method. Of the dynamic clustering models and methods that were specified in this dissertation, the dynamic clustering model proposed in Chapter 4 (MMVAR) offers the most comprehensive statistical model for describing individual dynamics. This model estimates the within-cluster variance of individuals' underlying coefficients, and combines describing individuals dynamics and clustering into a single statistical model. Compared to the MMVAR, the dynamic clustering model proposed in Chapter 3 (LCVAR) makes more restrictive assumptions of the underlying distributions. That is, the LCVAR model assumes within-cluster homogeneity in the individuals' true coefficients. The LCVAR model can consequently be estimated with greater simplicity and statistical power. This allows to fit models with a higher number of lags and outcome variables, as illustrated in our empirical example in Chapter 3 which estimated LCVAR models that included eight outcome variables over three lags.

Aside from differences in the underlying model assumption of the dynamic clustering procedures that were introduced in this dissertation, the distinct estimation procedures we proposed for these models and methods differed vastly, for instance, in their computation time. The estimation procedure proposed in Chapter 2 terminates after just two successive estimation steps: (1) describing individuals dynamics through a vector-autoregressive (VAR) model and (2) clustering individuals based on their VAR coefficients by calculating their probabilistic cluster membership. In contrast, the estimation procedures employed in Chapters 3 and 4 carry out these two steps repeatedly and until convergence, because they combine these two steps in their iterative estimation procedures, such as the expectation-maximization algorithm that was detailed in Chapter 3. While these iterative estimation procedures prevent a premature compression of the longitudinal data, they rely on a much higher number of estimation steps. This increases the computation time substantially, changing from the order of a few minutes to the order of a few hours or even days for common data sets.

We investigated the estimation performance of each of the three different dynamic clustering procedures that we proposed in this dissertation through accompanying simulation studies. In each simulation study we determined the recovery of cluster memberships and cluster parameters (henceforth denoted recovery). In Chapter 2 data for the simulation were generated with and without within-cluster variance of individuals' underlying VAR coefficients. For models containing four outcome variables over one lag, our proposed estimation method showed on average moderate recovery when the number of time-points was as low as 51 and the number of individuals was 30, 60, or 120. In Chapter 3 all data for the simulation were generated without within-cluster variance of individuals' underlying VAR coefficients, in line with the underlying assumption of the LCVAR model. On models that contained four outcome variables and up to two lags, our estimation procedure displayed very good recovery when the number of time-points equalled 50 and the number of individuals equalled 120. The data we generated in Chapter 4 had within-cluster variation in individuals' underlying VAR coefficients. For models with two outcome variables and one lag, 75 time-points and 100 individuals were sufficient for good recovery by our estimation procedure, as long as the distance between clusters was sufficiently large. The distance between the true underlying cluster parameters was consistently found an important factor in the recovery across all simulations. This between-cluster distance is relative to the within-cluster distance. In the simulation study in Chapter 2 we show that also the distance within clusters and the interaction between within-cluster and between-cluster distance has a large effect on the recovery.

Particularly the estimation of the dynamic clustering model proposed in Chapter 4 suffered from estimation and convergence problems when the number of parameters was high. Because this dynamic clustering model estimates the random effect (co-)variances of selected model parameters, it is especially for this model important to restrict the number of parameters, by keeping the number of outcome variables limited and by imposing constraints on the model. In Chapter 4 we discuss sensible constraints than can be imposed, such as constraining certain parameters to be equal across clusters. In the future, the potential of efficient estimation techniques such as Bayesian estimation and structural equation modelling could be explored for estimating extensive models, like the MMVAR, with a high number of parameters and in a time-efficient manner.

6.1.1 Implications of centering

In Chapter 4 we considered how the interpretation of the intercept coefficient in a multilevel VAR model is impacted by within-person centering and by not centering the predictors. A simulation study in Chapter 4 also provided insight into the bias of autoregressive and cross-regressive coefficients that occur in multilevel VAR models — and therefore in MMVAR models as well — when the predictors are either within-person centered or not centered. This simulation extended previous work on the bias of autoregressive coefficients (e.g., Hamaker & Grasman, 2015) to the multivariate (i.e., cross-regressive) case. We showed that within-person centering leads to bias in the auto-regressive coefficients, while not centering does not lead to bias in the auto-regressive coefficients. If the variance of the random VAR coefficients is non-zero, cross-regressive coefficients are estimated with bias for both within-person centered and not centered predictors. The bias of the cross-regressive coefficients is more severe in multilevel VAR models with uncentered predictors rather than within-person centered predictors. We showed that all these biases reduce with an increasing number of time-points.

6.2 Empirical applicability

Each dynamic clustering procedure that was proposed in this dissertation, was illustrated through a concrete empirical example. We showed how the developed models can be used to enhance the quality of psychological research. For example, in Chapter 4 we show that the exploratory nature of clustering procedures can result in more justified classifications than the assignment of individuals into groups based on observed variables. In dynamic clustering, individuals are classified based on the characteristic of interest — in this case their emotion dynamics — instead of based on an observed variable. Consequently, individuals who present atypical dynamics given their observed variables, are done more justice in an exploratory classification where they are grouped directly based on the dynamics of interest. On top of that, identifying such 'atypical' individuals through exploratory clustering is often of great interest to empirical researchers. In Chapter 4, for instance, we identified a few individuals who displayed atypical emotion dynamics for their age group; in follow-up analyses it would be interesting to find out why these individuals displayed such atypical emotion dynamics.

Dynamic clustering can facilitate the personalising of feedback and interventions to be tailored to the individual. Because dynamic clustering constitutes a solution to analyse ecological momentary assessment data in a way that accounts for atypical cases and between-individual heterogeneity, while evading common issues of single-individual models, such as overfitting (Bulteel et al., 2018) and low statistical power (Mansueto et al., 2020). Particularly the dynamic clustering model proposed in Chapter 4 could be useful for making generalisations, because it provides estimates of within-cluster variation, which can be compared to the between-cluster variation. This can give an indication of the usefulness of the clustering to summarise the dynamics and to provide generalisations. Dynamic clustering thus accounts for atypical cases and between-individual heterogeneity, while allowing for generalisations to the population of individuals. This is unlike a model that is entirely focused on the individual, as is the case in single-individual models.

6.3 Avenues for future research

6.3.1 Extensions to complex data structures

In Chapter 5 we proposed a comprehensive model framework that encompasses various longitudinal models that account for between-individual differences to some extent. The dynamic clustering models that are introduced in Chapters 3 and 4 can be connected to other longitudinal models through the proposed model framework. In Chapter 5 we discuss how this comprehensive model framework can be extended further. This discussion therefore gives a road-map to extending the dynamic clustering models that have been proposed in Chapters 3 and 4 as well. For instance, currently our dynamic clustering models contain clusters of individuals at the between-individual level, additionally clusters of time-points at the within-individual level could be included. Alternatively, the clusters of individuals could arise from a number of different nesting units, for instance, individuals could be sampled from different schools or different countries. Then the mixing proportions that indicate with which probability members of certain clusters occur might differ across the different nesting units. In that case, the model framework would need to be extended to account for the nesting structure of individuals.

Another possible extension of our model framework concerns the inclusion of the social context into the longitudinal model, for instance by incorporating a dyadic structure. The current within-individual approaches, such as the VAR model, have been criticised for not taking the social context sufficiently into account when modelling experiences and emotion dynamics (Dejonckheere et al., 2020; Lapate & Heller, 2020; Mestdagh & Dejonckheere, 2021). After all, psychology understands humans to be social beings, who often emote together, and impact and regulate one another (Butler, 2011). In order to include the social context into ecological momentary assessment studies, empirical researchers have begun to study emotional dynamics in dyadic relationships, for instance between romantic partners, or parent and child. In the future it would be useful to develop appropriate dynamic clustering procedures that shed light on between-dyadic differences in interpersonal influences. This will provide insight into how people influence each other's emotion dynamics and how people differ in influencing each other's emotion dynamics. Such dyadic dynamic clustering extensions could reveal how dynamic processes arise in interpersonal contexts, such as romantic relationships, supervisory relationships or families. The resulting dynamic clustering procedures might illuminate research questions on interpersonal processes of emotional connection and exchange between people.

6.3.2 Robustness of the dynamic clustering procedures

The dynamic clustering models and methods developed in Chapters 2, 3, and 4, as well as the models that are unified into a common model framework in Chapter 5, rely on a number of different statistical assumptions. Most notably, they make assumptions that certain distributions are Gaussian or that various parameters are mutually independent. It is valuable to study the extent to which these models are robust against violations of these assumptions. Furthermore, variants of these models that do not rely on these assumptions could be developed. The Bayesian framework offers interesting possibilities for this avenue. For instance, the dynamic clustering model proposed in Chapter 4 assumes the random effects within a cluster to follow a Gaussian distribution, by employing the Bayesian framework, this assumption could be relaxed.

Investigating the robustness to assumption violations could be particularly useful for the dynamic clustering model proposed in Chapter 3. As discussed in the section on model specification and estimation above, the dynamic clustering model proposed in Chapter 3 can be seen as a special instance of the dynamic clustering model proposed in Chapter 4, with an additional assumption. Specifically, the model assumes within-cluster homogeneity. The Chapter 3 model has the benefits of (1) faster estimation and (2) higher statistical power. As a result, more complex models with a higher number of variables and lags can be estimated with the Chapter 3 model. It would be interesting to find out how robust the Chapter 3 model is to violations of the additional assumption, in order to determine in which instances the Chapter 3 model should be preferred over the Chapter 4 model and vice versa.