Dynamic clustering

Ernst, Anja Franziska

DOI:
10.33612/diss.196176258

Chapter 1

Introduction

In the social sciences an elementary part of research over the past decades was to study how individuals' emotions and behaviours change over time. Nowadays, smartphones and tablets are enabling researchers to collect time-intensive, multivariate data through ecological momentary assessment. During ecological momentary assessment people are asked to report multiple times a day on, for example, their emotions and experiences, for one or more weeks. The availability of the resulting intensive longitudinal data has brought about a shift towards studying within-individual processes. By transitioning away from longitudinal data that contains only few time-points towards intensive longitudinal data, the focus of the analyses shifts from describing growth and decline, towards describing the dynamics with which variables fluctuate and interact over time. In particular, social scientists are interested in the dynamics with which the emotions (e.g., happy mood), clinical symptoms (e.g., suicidal thoughts), and trans-diagnostic factors (e.g., rumination) of a single individual interact over time. These dynamics reflect, for instance, the propensity of emotions and experiences to carry on and persist over time, and the augmentation and blunting processes between different emotions or experiences. In recent years the social sciences have profited from the inclusion of such temporal dynamics into theories and assessment, particularly concerning the developments of emotions and psychopathology over time.

To analyse such intensive longitudinal data, particularly vector-autoregressive (VAR) models (Lütkepohl, 2005) have been employed. VAR models express within-individual fluctuation by describing how the subsequent system state relates to the current system state. By describing the dynamic interplay between variables over time, VAR models can provide insight into, for instance, the emotion dynamics of an individual. The coefficients of an individual's VAR model can then reveal the persistence of emotions over time, and the augmentation and blunting processes between different emotions. Such VAR models are developed for modelling the dynamics of a single individual over time. However, the interest often lies in modelling a group of individuals, studying not only the within-individual dynamics but also the between-individual differences.

In psychology, VAR models are often applied to the longitudinal data of each person individually. Thereby each person in the sample is described by a different statistical model and this model focuses on the dynamics of that single individual. Such single-individual models are called for in psychology, because they offer the possibility to tailor interventions and therapy to the individual, and to offer personalised feedback (e.g., Fisher & Boswell, 2016). So far, however, individualised models still face statistical issues that have yet to be overcome. For example, single-individual VAR models are often overfitted. As a consequence, these overfitted models tend to inaccurately express the dynamics of the individual that the model was exclusively fitted to (Bulteel et al., 2018). Further, with the number of time-points that are feasible in clinical practice, single-individual VAR models do not have sufficient statistical power (i.e., with 75 – 100 time-points, see Mansueto et al., 2020). Also, the VAR models that are uniquely fitted to the dynamics of single individuals often differ structurally

(i.e., qualitatively) between individuals (Hamaker et al., 2016). These structural differences hamper comparisons between individuals' VAR models.

Often, researchers want to generalise their results to a population of individuals, and/or pool the data of multiple individuals to assess how an individual compares to others. To this end, a model that encompasses multiple individuals is needed. Ideally, such a model integrates information across individuals, while accounting for the between-individual heterogeneity in the within-individual processes.

To address the issues that have been outlined above, various multi-individual VAR extensions have been proposed to account for between-individual heterogeneity. These VAR model extensions integrate several individuals into the same model, while accounting for between-individual heterogeneity in the underlying dynamics in some way. One of these extensions is the multilevel VAR extension (Bringmann et al., 2013; Rovine & Walls, 2006), which accounts for continuous between-individual heterogeneity through random VAR coefficients. Random coefficients are specific for every individual and consist of a fixed effect, which represents the population average across individuals, and a random effect, which represents the individual's deviation from this average. Often, it is assumed that the random effects stem from a Gaussian distribution. Random VAR coefficients therefore pool the estimates across individuals, while accounting for quantitative inter-individual differences in the VAR coefficients. Random coefficients can capture the quantitative between-individual heterogeneity that is present in the dynamic processes of various individuals. However, researchers are often interested in capturing between-individual heterogeneity that is characterised by qualitatively different processes.

1.1 Dynamic clustering

Oftentimes researchers want to identify subgroups of individuals who display qualitative differences in their dynamic processes. This is needed in order to be able to generalize results to the population despite individuals exhibiting qualitatively distinct dynamics processes. Further, it can be needed to pool the data of several individuals into a common statistical model, despite their qualitative differences, because pooling data avoids overfitting and increases the statistical power of the model. This can be done by using an observed variable, such as gender or age group, to divide the sample into subgroups (Voelkle et al., 2014). Commonly, however, the subgroups of individuals who display qualitative differences and the individuals' membership in these subgroups are unknown. Then these unobserved subgroups need to be inferred from the data in an exploratory manner, using clustering. Clustering strives to divide the heterogeneous sample into subgroups that are homogeneous (Lubke & Muthén, 2005). The purpose of the clustering analysis is to identify the different subgroups, or clusters, within the data and to classify individuals into these clusters. By being a truly data-driven procedure, clustering can account for between-individual heterogeneity for which the reasons are unobserved: the heterogeneity cannot and need not be explained by an observed variable.

The aim of this thesis is to develop *dynamic clustering* procedures that account for qualitative between-individual differences in intensive longitudinal data. The resulting dynamic clustering procedures uncover clusters of individuals who exhibit qualitatively different dynamic processes in their longitudinal data. Thereby unknown subgroups of individuals with similar dynamics are identified. Dynamic clustering allows the information of several individuals to be pooled, while accounting for qualitative between-individual heterogeneity in the underlying dynamics. To achieve our dynamic clustering procedures we propose novel dynamic clustering models and methods. In this introductory chapter we refer to a dynamic clustering procedure as a dynamic clustering model when the two crucial steps, summarising

individuals' dynamics and clustering, are integrated into a holistic statistical model. Otherwise, we refer to a dynamic clustering procedure as a dynamic clustering method.

Our dynamic clustering procedures will be achieved by proposing multi-individual VAR model extensions that uncover clusters of individuals who exhibit qualitatively different dynamic processes. By basing our dynamic clustering procedures on the VAR model, the description of the average multivariate dynamic relations of the clusters will be in a format that researchers in the field are familiar with. Many empirical researchers know how VAR coefficients should be interpreted and these interpretations provide insight into the dynamic properties that are of interest to social scientists, such as the persistence, augmentation and blunting of emotions and experiences (Koval et al., 2012; Kuppens et al., 2010).

This dissertation introduces different dynamic clustering procedures and a unifying model framework for multi-individual longitudinal models. Our unifying model framework encompasses our dynamic clustering models and also many of the most prominent longitudinal models in the social sciences, biology, and medicine. Therefore, our model framework relates the dynamic clustering models that are proposed in this dissertation to other longitudinal models that are the current state of the art in the literature on dynamic modelling. We address two challenges for dynamic clustering: (1) accounting for continuous within-cluster heterogeneity in the VAR coefficients, and (2) adaptive estimation of the cluster-specific average VAR models. In the summary of Chapter 3 below we outline why an adaptive model estimation is advantageous.

1.2 Overview

Following this introductory chapter, the remainder of this thesis addresses the development, estimation, and model specification of dynamic clustering methods and models. Chapter 2 proposes a dynamic clustering method. Chapters 3 and 4 each propose a dynamic clustering model. Chapter 5 addresses the topic of model specification and model comparison by proposing a comprehensive model framework.

Chapter 2 describes a dynamic clustering method that proceeds in a two-step fashion. The method is estimated by (1) summarising each individuals' dynamics with a separate VAR model, and subsequently (2) clustering the individuals' VAR coefficients with a mixture model (McLachlan & Peel, 2004). By employing a mixture model to cluster VAR coefficients, the method allows for continuous within-cluster heterogeneity in the VAR coefficients. In a simulation study, the performance of our dynamic clustering method is compared to a dynamic clustering procedure that does not allow for within-cluster heterogeneity in the VAR coefficients (proposed by Bulteel et al., 2016). We show that when there is variation in VAR coefficients within the clusters, our dynamic clustering procedure outperforms the dynamic clustering procedure that does not allow for within-cluster heterogeneity. Both dynamic clustering procedures are illustrated using data that assessed the emotion dynamics of 366 individuals with ecological momentary assessment, as well as their depression and anxiety levels. We describe the resulting dynamic clusters in terms of the average emotion dynamics that are displayed in these clusters. We assign individuals into clusters based on their modal cluster membership probabilities to describe clusters in terms of their members' average scores on measures of depression and anxiety.

Chapter 3 presents a dynamic clustering model that is estimated in an adaptive manner. In contrast to the dynamic clustering method of Chapter 2, this model combines the two steps — describing individuals dynamics and clustering — into a holistic statistical model. Therefore, we refer to this dynamic clustering procedure as a model, rather than a method. A novel expectation-maximization algorithm is proposed that enables the *adaptive* estimation

of this dynamic clustering model, the resulting algorithm is implemented in an accompanying R function. Because the description of within-individual dynamics adapts during the cluster estimation, premature compression of the longitudinal data is prevented. The adaptive estimation enables a cluster-specific description of the within-individual dynamics where different clusters can be described by qualitatively different longitudinal models, for instance by VAR models with different numbers of lags. This dynamic clustering model assumes within-cluster homogeneity in the VAR coefficients. This means that VAR coefficients are equal for individuals in the same cluster, leading to no overlap in VAR coefficients between different clusters. The performance of the adaptive estimation procedure is assessed in a simulation study. We show that under perfect model specification, the recovery of cluster memberships and cluster parameters is excellent under a range of conditions, for instance, when the number of time-points is equal to 150. The dynamic clustering model is illustrated by an application to ecological momentary assessment data of 410 individuals; for each individual, numerous emotions were assessed at over 70 time-points over the course of a month. We describe the average emotion dynamics of the resulting clusters. Clusters are further described through their average scores on external measures of depression and anxiety when assigning individuals into clusters based on their modal cluster membership probabilities.

Chapter 4 introduces a dynamic clustering model that allows for within-cluster heterogeneity in the VAR coefficients and thereby extends the dynamic clustering model of Chapter 3. In this model, individuals' VAR coefficients are expressed as random coefficients where the fixed effects correspond to the mean coefficients in a given cluster, and the random effects represent the individual's deviation from the mean coefficients of the cluster they belong to. As a consequence, the variance of the random effects captures the heterogeneity within a given cluster. The dynamic clustering model of Chapter 4 can thus also be seen as an extension to the multilevel VAR model that adds the identification and assignment of individuals into clusters. Some of the key issues in model specification and estimation are considered. Particularly the issue of centering predictors is discussed and explored in a simulation study. We show that, in our model, centering predictors can lead to a bias in the VAR coefficients and that this bias vanishes with a high number of time-points. We propose an estimation procedure for the novel dynamic clustering model and assess its performance in a simulation study. The proposed model is applied to the ecological momentary assessment data of a heterogeneous sample, consisting of young and elderly adults. We describe the average emotion dynamics in the exploratory identified clusters and show that these clusters correspond to the distinct emotion dynamics that are commonly described for young and elderly adults. We also show that the exploratory assignment of individuals into clusters results in a more justified classification than the assignment of individuals into groups based on age. That is because individuals are classified based on the characteristic of interest — their emotion dynamics — instead of being classified based on an observed variable — such as age. By assigning individuals into clusters based on their modal cluster membership probabilities, we show that clusters differ significantly in their members' age, negative affect, and neuroticism.

Chapter 5 specifies a model framework that encompasses the dynamic clustering models proposed in Chapters 3 and 4, but also encompasses some of the most prominent longitudinal models in the social sciences, biology, and medicine, such as multilevel regression models (Skrondal & Rabe-Hesketh, 2008), growth curve models (McArdle, 1988), latent class growth analysis (Muthén & Asparouhov, 2008), and growth mixture models (Muthén & Asparouhov, 2008; Ram & Grimm, 2009). In this chapter the dynamic clustering models that are introduced in the dissertation are connected to other longitudinal models through the proposed model framework. The framework can be used to express various longitudinal models that aim to account for between-individual heterogeneity in different attributes, using separate approaches, across distinct disciplines and on dissimilar data structures. For example, the framework contains models that account for between-individual heterogeneity in

different attributes of the longitudinal data, such as the dynamics, the growth and decline, or the cyclical trends. By using distinct longitudinal models as concrete examples, the key characteristics of our model framework are discussed. Commonalities and differences of various models are highlighted and it is shown that a diverse set of models can be expressed in terms of our comprehensive framework. We outline recommendations for empirical researchers for model selection and model specification.

Chapter 6 summarises and compares the dynamic clustering procedures that were presented in this dissertation. We evaluate the estimation procedures we proposed for each dynamic clustering procedure by considering their differences in computation time, the number of parameters they can estimate, and the number of individuals and time-points they require to recover cluster memberships and cluster parameters well. We highlight the potential of dynamic clustering to enhance the quality of psychological research. Finally, avenues for future research are given. We focus particularly on extending our dynamic clustering procedures to accommodate more complex data structures, such as data with additional nesting structures or dyadic data. In the future, it might also be useful to investigate the extent to which our dynamic clustering procedures are robust to the violation of certain assumptions.