

## University of Groningen

### Dynamic clustering

Ernst, Anja Franziska

DOI:  
[10.33612/diss.196176258](https://doi.org/10.33612/diss.196176258)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2022

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Ernst, A. F. (2022). *Dynamic clustering: classifying people through ecological momentary assessment*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.  
<https://doi.org/10.33612/diss.196176258>

#### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

#### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

## Chapter 5

# A comprehensive model framework for between-individual differences in longitudinal data

### 5.0.1 Separating cross-sectional and longitudinal effects

In many fields of science longitudinal data on multiple subjects is used for answering research questions. For instance, for decades, an elementary part of research in the behavioural sciences has been to study how individuals' emotions and behaviours change over time. Multisubject longitudinal data illuminates heterogeneity between individuals in both their levels and their changes over time. The level is typically expressed by the within-person mean(s) for an individual, possibly at different points in a cyclical trend or at different points on a trajectory.<sup>1</sup> These within-person mean(s) are denoted as the individuals' *trend-cycles* (Hyndman & Athanasopoulos, 2018). The changes over time are assessed through the relationships between successive deviations from the trend-cycles at each time-point. These relationships express the individual *dynamics* (Hamaker et al., 2018; Jongerling et al., 2015).

Due to recent technological advances, smartphones and tablets are now providing an infrastructure that allows for the collection of time-intensive, multivariate data using ecological momentary assessment (Jebb et al., 2015). Intensive longitudinal data is loosely defined as a time-series with more than 20 time-points (Walls & Schafer, 2006). In fields like medicine (Burke et al., 2017), sociology (Tyler & Olson, 2018), psychology (Van der Krieke et al., 2016), and educational sciences (Blaauw et al., 2019; Malmberg & Martin, 2019), ecological momentary assessment is now commonly used. Therefore, in recent scientific literature a steep increase of publications based on intensive longitudinal data is visible (Hamaker & Wichers, 2017). Due to the rising collection of intensive longitudinal data, existing statistical models have been adapted and novel statistical models have been proposed to accommodate data of this length. When moving from longitudinal data that contains only few time-points towards intensive longitudinal data, the focus of the analysis typically shifts away from the coefficients that describe the trend-cycle, towards coefficients that describe the dynamics.

On top of studying individual trend-cycles and dynamics that are present in intensive longitudinal data, researchers are interested in describing their between-individual differences (Molenaar, 2004; Voelkle et al., 2014). In this paper, we discuss several ways to represent heterogeneity at the within-individual and the between-individual level for longitudinal data of various structures and of various lengths (i.e., panel data that includes only few time-points, and intensive longitudinal data). We do so by proposing a unifying general model

---

This chapter is currently submitted for publication as Ernst, A. F., Albers, C. J., & Timmerman, M. E. (Submitted-a). A comprehensive model framework for between-individual differences in longitudinal data.

<sup>1</sup>Cyclical trends describe, for instance, the average mood on mornings versus evenings. An example for a trajectory would be the average differences during first grade versus second and third grade (for trajectory trend-cycles, see e.g., Curran et al., 2010).

framework. At the within-individual level, we distinguish trend-cycles and dynamics represented in the longitudinal data. At the between-individual level, we account for discrete between-individual differences in trend-cycles and dynamics by identifying homogeneous sub-populations; we express additional continuous between-individual differences in these sub-populations through random coefficients.

The general model framework proposed in this paper encompasses several well-known longitudinal discrete-time models, including multilevel regression models (Skrondal & Rabe-Hesketh, 2008), growth curve models (McArdle, 1988), latent class growth analysis (Muthén & Asparouhov, 2008), growth mixture models (Muthén & Asparouhov, 2008; Ram & Grimm, 2009), vector-autoregressive models of lag order  $p$  (VAR( $p$ ), Lütkepohl, 2005), multilevel VAR models (Bringmann et al., 2013; Rovine & Walls, 2006), latent class VAR models (Anderlucci & Viroli, 2015; Ernst et al., 2020), and mixture multilevel VAR models (Ernst et al., Submitted-b). Many of these longitudinal models include a latent categorisation of individuals and/or random coefficients. Their use is widespread across disciplines. Due to differences in terminology, data structures and experimental set ups, the similarities and differences between these various models are not always eminently clear. This hampers a proper model selection and model interpretation. Therefore, this paper presents a unifying framework that represents these models in an accessible way. This will allow simple comparisons between the models and the various data structures, and ease their empirical application and interpretation.

In what follows, we first describe in which ways our model framework accounts for between-individual differences in longitudinal data. Second, we specify our general model framework and elaborate on the notation we employ throughout this manuscript. Third, by using famous longitudinal models as concrete examples, we discuss the key characteristics of our model framework. Fourth, we illustrate the characteristics of our model framework by briefly reviewing distinct longitudinal models that are encompassed in our framework. We highlight commonalities and differences of various models, and reflect upon important implications of the framework. Specifically, we show that despite vast differences in the data structures that these models were specified for, these models can all be united into a unified model framework. Fifth, we mention possible extensions to our model framework. Lastly, we conclude with recommendations for selecting and specifying longitudinal models that aim to account for between-individual differences.

## 5.1 Accounting for between-individual differences

Our model framework can be seen as a multilevel model where time-points (level 1) are nested within individuals (level 2). Level 1 thus corresponds to the within-individual level, level 2 to the between-individual level. In our model framework, within-individual differences are modelled at level 1 through a longitudinal model, accounting for the dynamics and trend-cycles; between-individual differences are modelled at level 2, in terms of the level 1 parameters, which describe the dynamics and trend-cycles. Between-individual differences are captured by discrete and/or continuous random variables (i.e., via latent categorisation and/or random coefficients, respectively). Random coefficients account for continuous between-individual differences, which emerge from a continuous distribution of deviations around a mean coefficient (De Leeuw & Meijer, 2008). As a consequence, random coefficients cannot account for discrete between-individual differences. To identify such discrete between-individual differences, latent categorisations are inferred from the data in an exploratory fashion (Lubke & Muthén, 2005). These exploratory models aim at uncovering latent sub-populations with discrete differences in trend-cycles and dynamics.

### 5.1.1 Latent categorisations

To account for unobserved sources of heterogeneity, many methods for identifying latent categorisations have been proposed, like for instance  $k$ -means or  $k$ -median clustering (Jain & Dubes, 1988; Lloyd, 1982). In this paper, we focus exclusively on model-based clustering methods, that is on latent class and finite mixture models (McLachlan & Basford, 1988; McLachlan & Peel, 2004; Vermunt & Magidson, 2003). Latent class and finite mixture models assume that the population comprises a finite number of *latent sub-populations* with distinct properties. These models offer a framework for formal inference, and provide a probabilistic assignment of individuals into latent sub-populations. We will denote this assignment as *latent categorisation* in what follows.

Instead of latent categorisations, random coefficients can be employed to account for continuous between-individual differences in model coefficients. In line with the literature, we denote a model involving categorisation into latent sub-populations a *latent class* model when the model does not allow for between-individual differences within the coefficients of a sub-population; in that case the sub-populations will be denoted as *latent classes*. Conversely, a model that allows within sub-population variability across individuals (i.e., by modelling the sub-population coefficients with random coefficients), will be denoted a *mixture model*, where the sub-populations are represented through so-called *components*. See, for instance, Muthén and Asparouhov (2008, Section 6.5.1), who denote a growth mixture model with no variance within latent categorisations (i.e., zero (co)variance for all random effects) as latent class growth analysis.

### 5.1.2 Random coefficients

While ideographic longitudinal models have been proposed for the analysis of intensive longitudinal data (Fisher & Boswell, 2016), it is often advantageous to pool information of several individuals while accounting for between-individual differences. This is commonly done through random coefficients. Individual random coefficients comprise a *fixed effect*, shared by all individuals, and a *random effect*, specific to one individual (e.g., Snijders & Bosker, 2011b). The fixed effect constitutes the population mean regression coefficient across individuals, while the random effect constitutes the individual's deviations from this mean coefficient. Random coefficients are predicted through a weighted average of the mean population coefficient and the individual's specific coefficient. An individual's random coefficient is therefore pulled (shrunken) towards the population mean coefficient (see De Leeuw & Meijer, 2008, Section 1.6.1). The model thereby accounts for continuous between-individual differences by allowing variation in the regression coefficients over individuals, whilst estimates from the population are pooled. Random coefficients for between-individual differences are often assumed to be multivariate normally distributed. Alternatively, Bayesian estimation methods have been proposed that allow for greater flexibility in the distribution of the random effects (e.g., Driver & Voelkle, 2018).

In a standard random coefficient model, the fixed effects correspond to mean coefficients of the entire population, and the random effects represent the individual's deviations from the population mean. If a mixture model is specified at the between-level, fixed effects correspond to mean coefficients of the mixture components, and the random effects represent the individual's deviations from the mean mixture component coefficients. Additional between-individual differences can thus be accounted for with random effects that represent the between-individual differences within the mixture component. In the special case of a mixture model involving only a single component, random effects represent the between-individual differences within this single component, which thereby corresponds to the complete population of individuals.

## 5.2 Model framework

We propose a model framework for multivariate intensive longitudinal data of multiple individuals, possibly including time-varying and/or time-constant covariates. Our framework includes a model part for both the trend-cycles and the dynamics. These two parts allow to disentangle change and stability represented in the time-series at the within-individual level. The model framework further includes two mixtures at the between-individual level, each with their associated number of components, where one mixture is for the trend-cycles and the other for the dynamics. These mixtures account for discrete between-individual differences, and thus aim to describe homogeneous sub-populations, which are latent (i.e., unobserved). Continuous between-individual differences within these latent sub-populations are expressed through random coefficients. The model thus has a within-individual part (at level 1), and a between-individual part (at level 2), with the two mixtures defined at level 2.

We underline random variables (i.e., the random coefficients and the expected value of an outcome variable) to distinguish them from fixed coefficients (see e.g., De Leeuw & Meijer, 2008). We use a tilde ( $\sim$ ) to indicate random effects, that is coefficients that are estimated with shrinkage. We will therefore refer to random coefficients as predictors rather than estimators (see De Leeuw & Meijer, 2008, Section 1.6.1). Coefficients that are not underlined represent fixed effects. Subscripts  $i$  and  $t$  indicate that the coefficient depends on individual  $i$  and time-point  $t$ ; superscripts of  $(k_1)$  and  $(k_2)$  indicate this individual's membership in a latent sub-population for the trend-cycles and the dynamics, respectively. Latent sub-population memberships stay constant for each individual across all time-points. We will employ the following model framework

$$\text{Level 1 : } \underline{y}_{it} = \underline{\mu}_{it}^{(k_1)} + \underline{\zeta}_{it}^{(k_2)} + \underline{\eta}_{it}^{(k_1)} \quad \underline{\eta}_{it}^{(k_1)} \sim N\left(0, \tau^2^{(k_1)}\right) \quad (5.1)$$

$$\underline{\mu}_{it}^{(k_1)} = f^{(k_1)}(\mathbf{x}_{it}, \mathbf{r}_{it}) \quad (5.2)$$

$$\underline{\zeta}_{it}^{(k_2)} = \left( \sum_{a=1}^{p^{(k_2)}} \underline{\zeta}_{i,t-a}^{(k_2)} \underline{\rho}_{ia}^{(k_2)} \right) + g^{(k_2)}(\mathbf{x}_{it}) + \underline{\epsilon}_{it}^{(k_2)} \quad \underline{\epsilon}_{it}^{(k_2)} \sim N\left(0, \sigma^2^{(k_2)}\right) \quad (5.3)$$

$$\text{Level 2 : } \underline{\rho}_{ia}^{(k_2)} = \rho_a^{(k_2)} + \underline{\tilde{\rho}}_{ia}^{(k_2)} \quad \underline{\tilde{\rho}}_{ia}^{(k_2)} \sim N\left(0, \omega_a^{(k_2)}\right). \quad (5.4)$$

- ◇  $\underline{y}_{it}$  represents the  $1 \times 1$  observation for individual  $i$  ( $i = 1, \dots, N$ ) at time-point  $t$  ( $t = 1, \dots, T_i$ ).<sup>2</sup> The model specified in Equations 5.1 – 5.4 contains only a single outcome variable in  $\underline{y}_{it}$ . Such an autoregressive (AR) model can be extended to a vector-autoregressive (VAR) model by considering a vector of outcome variables (Lütkepohl, 2005). While we include such multivariate models in our review, we use the univariate notation throughout this manuscript for simplicity of notation.<sup>3</sup> Whether a model discussed in this paper is specified for multiple outcome variables can be inferred from Tables 5.2a and 5.2b; in these tables we outline the data structure that the models were specified for. Multivariate models are advantageous in that they allow insight into both the covariance between temporal effects, and the co-variation between different variables.
- ◇  $\underline{\mu}_{it}^{(k_1)}$  represents the  $1 \times 1$  trend-cycle for individual  $i$  at time-point  $t$ . Every individual's trend-cycle is assumed to arise from one of  $K_1$  ( $k_1 = 1, \dots, K_1$ ) latent sub-populations. Depending on the structure of  $f^{(k_1)}$ ,  $\underline{\mu}_{it}^{(k_1)}$  may be represented through random coefficients. This implies then that  $\underline{\mu}_{it}^{(k_1)}$  is modelled at level 1 and at level 2, examples of

<sup>2</sup>We assume a pre-sample of  $p$  previous observations is available, as common in  $\text{AR}(p)$  notation.

<sup>3</sup>Similarly, we consistently write  $\mathbf{x}_{it}$  in boldface, even when it denotes a scalar.

this are shown in the second column of Table 5.1. In a multilevel VAR model,  $\underline{\mu}_{it}^{(k_1)}$  is represented as a random coefficient for the (conditional) within-person mean; in a growth curve model,  $\underline{\mu}_{it}^{(k_1)}$  is represented as a random coefficient for the trajectory, and is based on the random intercept coefficient and the random slope coefficient(s).

- ◇  $\underline{\zeta}_{it}^{(k_2)}$  represents the  $1 \times 1$  state of the individual at time-point  $t$  (de-trended from  $\underline{\mu}_{it}^{(k_1)}$ ); the model describing the autoregressive process under which  $\underline{\zeta}_{it}^{(k_2)}$  arises therefore describes the individual's dynamics. Every individual's dynamics are assumed to arise from one of  $K_2$  ( $k_2 = 1, \dots, K_2$ ) latent sub-populations.
- ◇  $\underline{\eta}_{it}^{(k_1)}$  represents no-carry-over residuals. These residuals do not carry over to the next time-point. Standard (V)AR models (see e.g., Lütkepohl, 2005) usually exclude no-carry-over residuals. In VAR models, failing to account for no-carry-over residuals can lead to biased estimates when measurement error is present (Schuurman et al., 2015).
- ◇ For individuals within latent sub-population  $k_1$ , no-carry-over residuals,  $\underline{\eta}_{it}^{(k_1)}$ , are distributed with variance  $\tau^2^{(k_1)}$  and mean zero.
- ◇ The function  $f^{(k_1)}$  describes the structure of the trend-cycle,  $\underline{\mu}_{it}^{(k_1)}$ , which is potentially time-varying. The structure of the function  $f^{(k_1)}$  may depend on latent sub-population  $k_1$ .  $\underline{\mu}_{it}^{(k_1)}$  can be estimated based on covariates that contain (functions of) time, collected in  $1 \times q_r$  design vector  $\mathbf{r}_{it}$ , and/or based on potentially time-varying covariates, collected in  $1 \times q_x$  design vector  $\mathbf{x}_{it}$ . An intercept term can be included in the function as well, this term will correspond to the within-person mean. For different specifications of  $f^{(k_1)}$ , see examples in Table 5.1. The trend-cycle element could, for instance, be modelled using a random coefficient model of the form  $f^{(k_1)} = \mathbf{x}_{it}\boldsymbol{\beta}_i^{(k_1)} + \mathbf{r}_{it}\boldsymbol{\gamma}_i^{(k_1)}$ . In this manuscript we distinguish in our notation covariates that contain (functions of) time,  $\mathbf{r}_{it}$ , and all other covariates,  $\mathbf{x}_{it}$ , to highlight their theoretical differences. To distinguish the coefficients associated with these different covariates, we denote  $\boldsymbol{\gamma}$  as the vector of coefficients associated with (functions of) time,  $\mathbf{r}_{it}$ , and we denote  $\boldsymbol{\beta}$  as the vector of coefficients that is associated with all other covariates,  $\mathbf{x}_{it}$ . An example for a covariate that is a (function of) time,  $\mathbf{r}_{it}$ , would be the school year of a child, like in the illustrative example of Young and Hunter (2015) that is included in our review below. A covariate that is not a function of time,  $\mathbf{x}_{it}$ , could be, for instance, the dosage of a medication at time  $t$ , as in the illustrative example of Funatogawa and Funatogawa (2012) that is described in our review section below.
- ◇  $\underline{\rho}_{ia}^{(k_2)}$  represents the random autoregressive coefficient of the  $\text{AR}(p^{(k_2)})$  model at lag  $a$  ( $a = 1, \dots, p^{(k_2)}$ ). Its value may depend on the  $k_2$  latent sub-population membership of individual  $i$ . This coefficient represents the direct influence of  $\underline{\zeta}_{i,t-a}^{(k_2)}$  on  $\underline{\zeta}_{it}^{(k_2)}$ , as opposed to the indirect influence more previous measurements may have through the lagged associations.
- ◇ Instead of including intercept terms and the potentially time-varying covariates, contained in  $\mathbf{x}_{it}$ , in the function  $f^{(k_1)}$ , they can be included in the function  $g^{(k_2)}$ . The structure of the function  $g^{(k_2)}$  may depend on latent sub-population  $k_2$ . Depending on the structure of  $g^{(k_2)}$ , this function may contain random coefficients. This implies then that this function is modelled at level 1 and at level 2, examples of this are shown in the tenth column of Table 5.1. The classic VAR model contains its' intercept in this function (Lütkepohl, 2005), as can be seen in Table 5.1; a VAR model that also contains covariates,  $\mathbf{x}_{it}$ , in function  $g^{(k_2)}$  is sometimes referred to as a VARX model (see Lütkepohl, 2005, Chapter 10). Unlike an intercept term included in  $f^{(k_1)}$ , which

represents an individual's within-person mean, VAR intercepts lack a directly meaningful interpretation. Covariates included in  $f^{(k_1)}$  influence only one measurement, and these influences do not carry over to any future time-point. In contrast, the influence of covariates included in  $g^{(k_2)}$  carry over to the next time-points via the autoregressive process. For different specifications of  $g^{(k_2)}$ , see examples in Table 5.1.

- ◇  $\underline{\varepsilon}_{it}^{(k_2)}$  represents white noise error. Just as in our model, state-space models (Hamilton, 1994) partition the deviations from a prediction into (1) white noise error,  $\underline{\varepsilon}_{it}^{(k_2)}$ , representing meaningful yet unpredictable fluctuations, and (2) no-carry-over residuals,  $\underline{\eta}_{it}^{(k_1)}$ , which are often taken to represent measurement error (Schuurman et al., 2015). The white noise error differs from the no-carry-over residuals in the following, once the unpredictable white noise error is observed, it is useful for predictions of future time-points, this is not the case for no-carry-over residuals.
- ◇ For individuals within latent sub-population  $k_2$  white noise errors,  $\underline{\varepsilon}_{it}^{(k_2)}$ , are distributed with variance  $\sigma^2^{(k_2)}$  and mean zero.
- ◇  $\rho_a^{(k_2)}$  represents the fixed autoregressive (AR) coefficient in latent sub-population  $k_2$  at lag  $a$ .
- ◇  $\tilde{\rho}_{ia}^{(k_2)}$  represents the random effect for individual  $i$  for the AR coefficient at lag  $a$ . Random effects in latent sub-population  $k_2$  are distributed with variance  $\omega_a^{(k_2)}$  and mean zero. When multiple random coefficients are included in a model (for instance in  $f^{(k_1)}$  or  $g^{(k_2)}$ ), it is common to model the covariance between all random effects included in the model with covariance matrix  $\Omega^{(k_1),(k_2)}$ , as is elaborated on below.

### 5.2.1 Assumptions

The model framework as specified here relies on the assumptions that individuals are independent, and that the no-carry-over-residuals, the white noise error, and the random effects all are uncorrelated with the covariates contained in  $\mathbf{x}_{it}$ , and  $\mathbf{r}_{it}$ . Further, the no-carry-over-residuals, the white noise error and the random effects are all assumed to be uncorrelated with each other. All independence and distributional assumptions are expressed conditional on the covariates. While the model framework accounts for trend-cycles and fluctuations caused by covariates, its model parameters are time-invariant. This implies that this model framework assumes stationarity of the time-series' auto-covariance. When the autoregressive process under which  $\underline{\varepsilon}_{it}^{(k_2)}$  arises is included in the model, time-points need to be equidistantly placed for all individuals. Whenever time is included as a predictor, as is done in trajectory models, it is necessary from an interpretational point of view that individuals are assessed on a common time-axis. When individuals are assessed on a common time-axis, the model framework can be extended to include, for example, time-point-specific residual variances (see e.g., Anderlucci & Viroli, 2015). We focus here on models for discrete time, extensions to continuous time models are possible but neglected in this paper (e.g., Driver & Voelkle, 2018; Oravecz et al., 2009).

Every individual is assumed to belong to one of  $K_1$  ( $k_1 = 1, \dots, K_1$ ) latent sub-populations, and to one of  $K_2$  ( $k_2 = 1, \dots, K_2$ ) latent sub-populations. All sub-population memberships are unobserved and have to be predicted based on the longitudinal data. Each of the latent sub-populations,  $k_1$  and  $k_2$ , thus correspond to a discrete model for the trend-cycle element and the dynamic element, respectively. They correspond to sub-populations of individuals, who on the one hand display similar within-person means, trajectories, and influences of covariates on the within-person mean (i.e., trend-cycles), and on the other hand a similar temporal correlation structure (i.e., dynamics).

All random effects<sup>4</sup> (i.e., all coefficients indicated with a tilde ( $\sim$ ) in our notation), are assumed to arise from a multivariate normal distribution with mean vector zero,  $N(\mathbf{0}, \mathbf{\Omega}^{(k_1), (k_2)})$ . In this paper we restrict ourselves to multivariate normally distributed random effects throughout, to keep the notation and the framework uniform. More relaxed distributional assumptions can be made for random effects, however, for instance when using Bayesian estimation (see e.g., Driver & Voelkle, 2018). When multiple random coefficients are included, it is common to model the covariance between all random effects included in the model, on top of modelling their variance. For instance, in the multilevel AR model proposed in Jongerling et al. (2015), random effects are included for the AR coefficient,  $\underline{\rho}_i$ , for the variance of the white noise error,  $\underline{\sigma}_i^2$ , and for an intercept term,  $\underline{\beta}_i$ , which in this specific model describes the within-person mean. All random effects are then assumed to arise from a multivariate normal distribution with the following covariance matrix,  $\mathbf{\Omega}$ ,

$$\begin{pmatrix} \underline{\tilde{\beta}}_i \\ \underline{\tilde{\rho}}_i \\ \underline{\tilde{\sigma}}_i^2 \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{\beta}^2 & & \\ \omega_{\beta, \rho}^2 & \omega_{\rho}^2 & \\ \omega_{\beta, \sigma^2}^2 & \omega_{\rho, \sigma^2}^2 & \omega_{\sigma^2}^2 \end{pmatrix} \right).$$

Particularly in growth curve models, which include a random coefficient associated with indicators for time,  $\mathbf{r}_{it}$ , it is important for the random intercept term, contained in  $\underline{\boldsymbol{\gamma}}_i^{(k_1)}$ , and the random slopes, also contained in  $\underline{\boldsymbol{\gamma}}_i^{(k_1)}$ , to not be specified as uncorrelated. This is so, because changes to the time scale, contained in  $\mathbf{r}_{it}$ , will cause changes in the magnitude of the correlation between the random intercept coefficient and the random slope coefficient (Skrondal & Rabe-Hesketh, 2008).

### 5.2.2 Elements of our model framework

Whenever at least one design vector is included in a model, we do not specify intercepts as separate terms anymore, but include them as a vector of ones in  $\mathbf{x}_{it}$  (or in  $\mathbf{r}_{it}$  when  $\mathbf{x}_{it}$  is not included in the model). In case no design vectors are specified, we simply represent the intercept terms as  $\beta$ . An intercept term may depend on a latent sub-population  $k_1$  ( $\beta^{(k_1)}$ ), or on a latent subpopulation  $k_2$  ( $\beta^{(k_2)}$ ). An intercept term may also be represented through a random coefficient ( $\underline{\beta}_i = \beta + \underline{\tilde{\beta}}_i$ ). We denote the vector of covariates that are associated with the fixed effects by the same letter as the vector of covariates that are associated with the random effects. In general, it is possible to include covariates in the fixed effects design vector that are not included in the random effects design vector; however, we will neglect this possibility here for simplicity of notation. Also for simplicity of notation, symbols are consistently bold or not bold throughout this manuscript. Whether a model is specified for multiple outcome variables can be seen in Tables 5.2a and 5.2b, but it is not reflected in the notation.

Equation 5.1 encompasses the trend-cycles and the dynamics of the time-series via  $\underline{\boldsymbol{\mu}}_{it}^{(k_1)}$  and  $\underline{\boldsymbol{\zeta}}_{it}^{(k_2)}$ , respectively. To illustrate the properties of the model framework parts, it is useful to first consider a standard random coefficient model and then to add elements. The formulas for the following models are displayed in Table 5.1. A classic multilevel longitudinal model, the random coefficient model, expresses the dependence between the observations of an individual by individual-specific random coefficients that are shared by the measurements of all time-points of this individual (Skrondal & Rabe-Hesketh, 2008). Linear trajectories, such as growth or decline, can be accounted for by adding (functions of) time,  $\mathbf{r}_{it}$ , as covariate(s) to the random coefficient model; this results, for instance, in a growth curve model as displayed in Table 5.1 (McArdle, 1988).

<sup>4</sup>In line with the definition we gave above, random effects are coefficients that are estimated with shrinkage.



When measurements of the same person that are closer spaced in time are assumed to be more similar, one can add to the random coefficient model lags of the outcome variable as a covariate; then the associated coefficient,  $\rho_a^{(k_2)}$ , represents the autocorrelation of time lag  $a$ . In the multilevel literature, a model that includes lags of the response variable as a covariate is commonly called a *lagged response model* (Skrondal & Rabe-Hesketh, 2008) or a prospective change model (Larson & Almeida, 1999). Caution should be exercised when adding a lagged outcome variable as a covariate to a random coefficient model: Including the lagged outcome variable as a covariate in a model that contains random effects, will render the initial outcome endogenous, because the assumption of no association between the random effects and the covariates is violated (Rabe-Hesketh & Skrondal, 2013; Skrondal & Rabe-Hesketh, 2014). Also, when this AR coefficient is modelled with random effects, within-person centering will lead to a downward bias in the random AR coefficient,  $\rho_{i a}^{(k_2)}$ , (Hamaker & Grasman, 2015). This bias will vanish, however, with a high number of time-points (Asparouhov et al., 2018; Ernst et al., Submitted-b).

Alternatively to including lags of the outcome variable as a covariate, models have been employed that allow a type of residuals to be auto-correlated, as in the models proposed by Wang et al. (2012) and Anderlucci and Viroli (2015) that are reviewed below. These models are commonly denoted *lagged residual models* in the multilevel literature (Skrondal & Rabe-Hesketh, 2008). Our framework notation can offer a nice illustration of the difference between the lagged response model and the lagged residual model (for both see Skrondal & Rabe-Hesketh, 2008): When a model contains a lagged term, and the intercept and the potential covariates are contained in  $g^{(k_2)}$ , then the model is a lagged response model; when a model contains a lagged term, and the intercept and the potential covariates are contained in  $f^{(k_1)}$ , then the model is a lagged residual model. For models that include lagged outcome variables or lagged residuals, the covariance structure of the outcome variable is model implied (see Skrondal & Rabe-Hesketh, 2004, for an introduction to the autoregressive response and the autoregressive residual structure). The multilevel framework allows, thus, to model three distinct types of temporal dependency: (1) persistent dependency via shared individual-specific random coefficients, (2) immediate dependency via time-varying covariates, and (3) intermediate dependency that lingers for some time and decreases as time goes by via AR coefficients.

(V)AR coefficients,  $\rho_a^{(k_2)}$ , can be modelled as fixed coefficients (e.g., Anderlucci & Viroli, 2015) or as random coefficients (e.g., Bringmann et al., 2013; Cho et al., 2018; Driver & Voelkle, 2018; Jongerling et al., 2015; Schuurman et al., 2016). In this manuscript we denote models that represent (V)AR coefficients,  $\rho_a^{(k_2)}$ , through random coefficients as multilevel (V)AR models (Bringmann et al., 2013; Rovine & Walls, 2006).

The traditional autoregressive models account for the temporal dependencies through fixed coefficients (see e.g., Skrondal & Rabe-Hesketh, 2008). In these models, AR coefficients,  $\rho_a^{(k_2)}$ , are often included to "adjust" for the previously observed value of the outcome variable, not because the researchers are inherently interested in the autoregressive properties of the longitudinal data (e.g., Funatogawa & Funatogawa, 2012). In such models the inclusion of AR coefficients is often guided by statistical considerations, rather than substantive interest (see Snijders & Bosker, 2011a). In that case, the AR coefficients are treated as nuisance parameters. That is, they are included to obtain a good model fit but the coefficients are of secondary interest. In contrast, recent psychological research on emotions has focused particularly on the AR coefficients as interesting parameters in themselves. For this, autoregressive and potentially cross-regressive properties are modelled using (V)AR models (Lütkepohl, 2005). In a multivariate VAR model of emotion ratings, the cross-regressive coefficients can shed light on augmentation and blunting processes between different emotions. AR coefficients can provide insight into the tendency of specific emotional states to linger over time, a

concept called emotional inertia (Kuppens & Verduyn, 2017). Using multilevel (V)AR models, auto- and cross-regressive coefficients (i.e., VAR coefficients,  $\rho_a^{(k_2)}$ ) have been modelled as random coefficients to study between-individual differences in emotional inertia, augmentation, and blunting (Bringmann et al., 2013; Krone et al., 2018; Rovine & Walls, 2006). Such research has consistently linked between-individual differences in augmentation, blunting, and especially emotional inertia to psychological well-being and psychopathology (Houben et al., 2015; Koval & Kuppens, 2012; Kuppens et al., 2010; Kuppens & Verduyn, 2017; Van der Krieke et al., 2017; Van Roekel et al., 2018). In these multilevel (V)AR models, the random coefficients of the model are used to account for between-individual differences in (amongst others) the autoregressive properties, while the multilevel structure of the model can be exploited to relate these between-individual differences to between-level variables. For example, Koval et al. (2012) used multilevel models to relate level 2 between-person means on rumination to the random effects, which represent between-individual differences in trend-cycles and autoregressions of negative emotions. Multilevel (V)AR models have become widely used in the social sciences and many extensions to the model have been proposed. For instance, Haan-Rietdijk et al. (2016) extended the multilevel AR model to a state dependent (i.e., threshold-dependent) model; Oravecz et al. (2009) proposed a continuous time extension of the multilevel AR model.

### 5.3 Review of models

To illustrate our framework, below we briefly review multisubject models that contain latent categorisations and/or random effects at the between-level, with a particular interest for models that combine these two properties. As we will show, these models are special cases of our model framework. We focus here on models that are suited for longitudinal data of various structures (i.e., trajectory data or dynamic data) and various lengths (i.e., panel data that includes only few time-points or intensive longitudinal data). We restrict the review to latent class and mixture models (McLachlan & Basford, 1988; McLachlan & Peel, 2004; Vermunt & Magidson, 2003); other methods to infer latent categorisations from data, like  $k$ -means or  $k$ -median clustering (Jain & Dubes, 1988; Lloyd, 1982), are not considered here. We highlight that our review of models does neither constitute a systematic nor a comprehensive review. Rather, we selected the models that we discuss below, because they fit well into our model framework and because this selection of models, when presented together, offers a nice and complete illustration of all the elements that are presented in our model framework. For the most important models we discuss, the translation of the model into our framework notation is shown in Table 5.1, and the data structures the models were specified for are summarised in Tables 5.2a and 5.2b. We selected models to be included in these tables because they are either (1) prominent longitudinal models that we found instructive to illustrate in more detail, these models are listed under ‘standard models’ in the tables, or (2) models that are discussed in detail in our model review, these models are listed under ‘models included in the review’ in the tables. Table 5.1 links these models to our framework in Equations 5.1 – 5.4, and Tables 5.2a and 5.2b give pointers on how to use them, for instance, with which software these models can be estimated.

#### 5.3.1 Cross-sectional and random coefficient models

DeSarbo and Cron (1988) proposed a regression model for cross-sectional data, where between-individual differences in the regression slopes and intercepts are captured through one latent class model at the between-level. Lenk and DeSarbo (1999) extended this model to a mixture regression model by including random coefficients at the between-level; herewith they

captured the subject's deviation within the mixture components through random effects. As a result, this model can accommodate longitudinal data. The model by Lenk and DeSarbo (1999) is a random coefficient model, as defined in Table 5.1, that additionally includes a latent categorisation of individuals. The models proposed by DeSarbo and Cron (1988) and Lenk and DeSarbo (1999) can be extended to model trajectories by including (functions of) time,  $r_{it}$ , as covariates. Specifically, by including (functions of) time as covariates, these models can be extended to latent class growth analysis and growth mixture modelling, respectively, as illustrated below.

### 5.3.2 Trajectory models

By including coefficients for the trajectories, the random coefficient model as shown in Table 5.1, can be extended to a linear growth curve model (McArdle, 1988). Growth curve models allow for continuous differences in the trajectory coefficients, using random coefficients at the between-level (McArdle, 1988). Growth curve models are also referred to as latent growth curve models, latent growth models, and latent curve analysis (Frey, 2018). When the between-individual differences in the model coefficients are captured through discrete latent classes and not through continuous random coefficients, the resulting trajectory model is referred to as a latent class growth analysis (Muthén & Asparouhov, 2008). By modelling individual deviations from the latent sub-population mean via random coefficients, the latent class growth analysis can be extended to the growth mixture model (Muthén & Asparouhov, 2008). In a growth mixture model, every component is described by a growth curve model with continuous deviations from the component-specific fixed coefficients. Growth mixture models therefore account for continuous and discrete differences in the model coefficients, through random coefficients and mixture components, respectively (Muthén & Shedden, 1999). When the sub-populations are not latent but observed, growth mixture models are sometimes referred to as multiple group growth models (Kim et al., 2014; Muthén & Muthén, 2013).

Serang et al. (2016) proposed an estimation procedure for a non-linear growth mixture model, where in each mixture component the growth curve is modelled through random coefficients, to allow individuals of the same component to deviate from one another. The authors propose for their non-linear growth mixture model an estimation procedure in the Bayesian framework, and an estimation procedure in the frequentist framework, using structural equation modelling (SEM). In the SEM framework, the non-linear model is approximated through linearization using Taylor series expansion. In Table 5.1 below we present the linear approximation of their non-linear model, that is  $r_{it}$  contains partial derivatives of the growth function with respect to the fixed effects. In the SEM framework, individuals are often assumed to be assessed at equal time-points. The SEM framework approximation proposed by Serang et al. (2016) allows individuals to be assessed at different time-points. This acknowledges possible heterogeneity in the timing of assessments. For instance, some students would be assessed at the beginning of first grade, while others are assessed at the end of first grade. They applied their non-linear model to describe growth curves of reading ability in young children.

Young and Hunter (2015) proposed another growth mixture model: their model includes time-varying covariates via fixed component-specific coefficients, and random effects around the component-specific growth curve. The model allows for growth trajectories with varying lengths across participants. Young and Hunter (2015) employed the model to show that in a population of infants, subgroups exist which display distinct patterns of habituation during cognitive tasks.

In biostatistics and bioinformatics, a more complex data structure than the one we considered above is usually employed in longitudinal studies: genes are assessed on several

tissues and on each tissue, repeated measures are made of the gene at different time-points (e.g., microarray experimental designs: Ng et al., 2006; Wang et al., 2012). On each tissue the same genes are measured. In this data structure genes correspond to the individuals. A model for such data has been proposed by Ng et al. (2006). We consider in Table 5.1 a special case of their model where the number of tissues equals one; this model specification was used by Ng et al. (2006) as an illustrative example to study yeast cells. In the general model by Ng et al. (2006), every gene is assumed to arise from one of several latent sub-populations, and gene-specific random effects account for variation within these latent sub-populations. These gene-specific random effects account for the similarities that are shared over all time-points for a given gene on a given tissue<sup>5</sup> (i.e.,  $\tilde{\beta}_i^{(k_1)}$  in Table 5.1). Our model framework, which we have introduced above, accounts for differences between latent sub-populations only through fixed effects, while the differences within a latent sub-population are accounted for with random effects. The model proposed by Ng et al. (2006) models the differences between latent sub-populations through fixed and random effects, which induces dependency between genes from the same latent sub-population. Specifically, random effects are included for the time-points. These random time-point effects differ between different latent sub-populations, but are the same for genes from the same latent sub-population (i.e.,  $\tilde{\beta}_t^{(k_1)}$  in Table 5.1). Thus, these random time-point effects induce dependency between observations from the same time-points within a latent sub-population. In the case of such dependency, a conditional estimation procedure is necessary to estimate this model (i.e., conditional on the between latent sub-population random effects,  $\tilde{\beta}_t^{(k_1)}$  in Table 5.1). The inclusion of such random temporal effects is only sensible when subjects (i.e., genes) are assessed on a common time-axis. Ng et al. (2006) employed the model as specified in Table 5.1 to identify 12 mixture components of yeast genes that exhibit distinct periodic expression patterns over the cell cycle. Their fixed effects,  $r_{it}\gamma^{(k_1)}$ , include a Fourier series approximation to account for periodic trends in the genes cell cycle (Spellman et al., 1998).

To allow for additional time dependence, the model by Ng et al. (2006) was extended by Wang et al. (2012), who include an autoregressive structure for the gene effects over one time lag (i.e., it is a lagged residual model, as defined above, with an autoregressive residual variance structure as described in Skrondal & Rabe-Hesketh, 2004). Thus, in the model by Wang et al. (2012) repeated measurements of the same gene are assumed to exhibit a temporal dependency that lingers for some time and decreases as time goes by, which is accounted for with an autoregressive residual structure.

### 5.3.3 Models that account for autoregressive dependence

Wang et al. (2012) extended the model by Ng et al. (2006) that was described above, by including an AR residual structure in place of the gene-specific effects. Their model, as shown in Table 5.1, thus includes (1) fixed effects that model periodic trends due to the genes cell cycle using a Fourier series approximation ( $r_{it}\gamma^{(k_1)}$ ), (2) time-point-specific random effects that are shared among the genes in the same latent sub-population ( $\tilde{\beta}_t^{(k_1)}$ ), and (3) an autoregressive residual structure (for  $\zeta_i^{(k_2)}$ ) that accounts for autoregression in the repeated measurements of a gene on a given tissue through fixed AR coefficients ( $\rho_1^{(k_2)}$ ). Gene effects are thus modelled through fixed AR coefficients,  $\rho_1^{(k_2)}$ , that differ across latent sub-populations but have no within latent sub-population variance. The gene effects,  $\zeta_i^{(k_2)}$  (containing  $\zeta_{it}^{(k_2)}$  at

<sup>5</sup>In the case more than a single tissue is included, this gene-specific random effect represents tissue effects and is of a length equal to the number of tissues considered (i.e., it contains a different coefficient for each tissue).

all  $T_i$  time-points), are described with an autoregressive residual variance structure given by

$$\frac{\sigma^2(k_2)}{1 - \rho_1^{(k_2)}} \begin{pmatrix} 1 & \rho_1^{(k_2)} & \dots & \rho_1^{T_i-1(k_2)} \\ \rho_1^{(k_2)} & 1 & \dots & \dots \\ \dots & \dots & 1 & \dots \\ \rho_1^{T_i-1(k_2)} & \rho_1^{T_i-2(k_2)} & \dots & 1 \end{pmatrix}$$

(cf., the autoregressive residual variance structure in Skrondal & Rabe-Hesketh, 2004). Wang et al. (2012) employed their model, as described in Table 5.1, to describe the yeast cell cycle of 237 genes (in this example, the number of tissues was again equal to one). They describe four latent sub-populations with distinct periodic expression patterns (i.e., trend cycles) and with distinct autoregressive properties. This extends the model by Ng et al. (2006), which identifies latent sub-populations with only distinct periodic expression patterns.

In medicine, Funatogawa and Funatogawa (2012) proposed a univariate lagged response model, to model the repeated administration of medication with dose-modification. In this model, the dosage of a medication is the covariate and included in the design vector  $\mathbf{x}_{i,t}$ . As discussed above, such time-varying covariates represent immediate influences on the outcome variable. The model uses fixed AR coefficients to account for the influence of past dosages that linger for some time and decrease as time goes by. The model does not include latent classes and accounts for between-individual differences only through random effects for the time-varying covariates. The model by Funatogawa and Funatogawa (2012) includes white noise error, and no-carry-over residuals,  $\eta_{i,t}$ , to account for measurement error. The model can thus be denoted a state-space model (Hamilton, 1994). As an illustration of the approach, Funatogawa and Funatogawa (2012) analysed clinical trial data for multiple sclerosis patients to estimate the dose–response relationship for each patient; the corresponding model is shown in Table 5.1. Their model and estimation procedure can handle unequal spacing of measurements by assuming a small unit of time and treating measurement occasions as missing.

### 5.3.4 Models with random autoregressive coefficients

Traditionally, growth curve modelling has been the dominant method to analyse longitudinal data in the social sciences. In recent longitudinal research, the focus has shifted from modelling trend-cycles, reflecting developmental processes over time, to describing the dynamics of more stable processes (Hamaker & Wichers, 2017; Jebb et al., 2015; Molenaar, 2004). For this purpose, (V)AR models have become popular, which express the dynamics via the (V)AR coefficients. The recent shift towards intensive longitudinal data resulted also in the development of models that express between-individual differences in the (V)AR coefficients,  $\rho_{i,a}^{(k_2)}$ , through random effects. Such a model can be seen as a multilevel extension to the standard (V)AR model (Lütkepohl, 2005). This extension is referred to as a multilevel (V)AR model (e.g., Bringmann et al., 2013; McNeish & Hamaker, 2020), mixed-effects (V)AR model (e.g., Gorrostieta et al., 2012) or dynamic multilevel model (Jongerling et al., 2015). Illustrations of the multilevel VAR model can be found in Rovine and Walls (2006), Schuurman et al. (2016) and Bringmann et al. (2013).

Multilevel (V)AR models extend the standard single-subject (V)AR model to a multisubject model. It does so by including random effects at the between-individual level for both the (V)AR coefficients and the trend-cycle coefficients. Jongerling et al. (2015) illustrate the importance of including additionally random effects for the white noise error variance,  $\sigma^2$ , to avoid biased estimates. Jongerling et al. (2015) highlight the ability of their multilevel AR model to capture dynamics and trend-cycles, or as they call it state and trait effects. In their model, trend-cycles, modelled by  $\beta + \underline{\hat{\beta}}_i$  in Table 5.1, represent the within-person mean. As

discussed above in the section entitled ‘Assumptions’, their multilevel AR model estimates all random effects included in their model jointly and thereby accounts for associations (i.e., co-variations) between the random effects. Jongerling et al. (2015) used their multilevel AR model to describe the between-individual differences in the carry-over effects (i.e., the AR coefficients) in a sample of women. Based on the distribution of random effects around the fixed AR coefficient they found that for a majority of woman, positive affect carries over from the previous to the next day.

Once we move from a univariate multilevel AR model to a multivariate multilevel VAR model, it becomes more difficult to account for the co-variations between random effects. This is so because the number of random effects increases exponentially with the number of outcome variables. Bringmann et al. (2013) proposed a multilevel VAR model that was estimated through fitting separate multilevel AR models for each outcome variable. The variances of the random effects are thus estimated directly in these univariate models. In contrast, the remaining covariances, that is the covariances between random effects of the different univariate models, have to be estimated in a second step, by calculating the covariance between the estimated random effects. Similarly, the correlation between the white noise error variances of the univariate models are estimated in a second step by calculating the correlations between the residuals of the different univariate models. Bringmann et al. (2013) employed their multilevel VAR model to show how individuals’ scores on different mood items dynamically interact over time.

Gorrostieta et al. (2012) proposed a multilevel VAR model to determine brain connectivity across multiple regions of interest in multisubject neuroimaging data. In their data structure, the multiple outcome variables correspond to the multiple regions of interest. They model all random effects included in their model (i.e., every element in the VAR coefficient matrix  $\underline{\tilde{\rho}}_i$ ) as mutually independent random variables. Thus, similarly to the estimation procedure employed by Bringmann et al. (2013), Gorrostieta et al. (2012) estimate marginal multilevel AR models for each outcome variable. In their model, the VAR coefficients depend on the condition at every time-point. In Table 5.1, we denote the condition of individual  $i$  at a time-point  $t$  by  $c_{it}$ . We refrain from including this as  $k_2$  in Table 5.1 for the following reason: unlike the latent variable,  $k_2$ , that varies at the between-individual level, the condition,  $c_{it}$ , is observed and varies with the time-points and individuals (i.e., within-individuals and between-individuals).

The multilevel VAR model proposed by Driver and Voelkle (2018) is a continuous time model, and thereby allows for non-equidistant and/or non-discrete measurement occasions. In Table 5.1 we present their model translated to discrete time. The Bayesian estimation procedure employed in their paper offers the advantage that a full random effects model can be achieved where every parameter is allowed to vary across individuals. Further, Bayesian estimation enables estimating a joint distribution for all parameters (i.e., simultaneous estimation), rather than estimating marginal models and then estimating the joint distribution step-wise, as done by Bringmann et al. (2013) and Gorrostieta et al. (2012). The model proposed by Driver and Voelkle (2018) can further include random coefficients for time-varying and time-constant covariates in  $g^{(k_2)}$ . The application in their paper, however, focuses on the implementation of a manifest intercept, corresponding to a random intercept in  $f^{(k_1)}$ ; in Table 5.1 we therefore present the latter form of the model. In their application, Driver and Voelkle (2018) employ their multilevel VAR model to describe the dynamics of overall satisfaction and health satisfaction, and the extent to which between-individual variation in such dynamics exists. Their sample consisted of 200 individuals each assessed on 29 time-points. Their analysis provided them also with the correlations between the random effects, which can be used to understand the between-individual differences in the dynamics. In their paper they highlight the advantages of the multilevel VAR model: by employing random VAR coefficients to describe the dynamics, the estimates of an individual’s VAR coefficients are

pooled with the data of other people, while between-individual differences are still accounted for.

### 5.3.5 Models that capture between-individual differences in autoregressive dependence through latent categorisations

The multilevel (V)AR models reviewed in the previous section accounted for between-individual differences exclusively via continuous latent variables (i.e., random effects). Alternatively, categorical latent variables can be used (i.e., latent classes or mixture components) to account for between-individual differences in VAR models.

Anderlucci and Viroli (2015) and Ernst et al. (2020) extended the standard VAR model to include latent classes at the between-level. In contrast to our model framework, which includes two mixtures at the between-individual level, these models employ a joint latent categorisation for the trend-cycle and the dynamics (i.e.,  $k_1 = k_2$  for all individuals). The models proposed by Anderlucci and Viroli (2015) and Ernst et al. (2020) thus assume that latent categorisations consist of individuals with both similar trend-cycles and similar dynamics. The models proposed by Anderlucci and Viroli (2015) and Ernst et al. (2020) do not include random effects, assuming thereby identical coefficients for all individuals belonging to the same latent class.

In their model, Anderlucci and Viroli (2015) express the covariance pattern of the multivariate longitudinal data via a Kronecker product between a  $m \times m$  covariance matrix  $\Sigma_{within}^{(k_2)}$  and a  $T \times T$  covariance matrix  $\Sigma_{between}^{(k_2)}$ , which represent the covariance matrices between  $m$  outcome variables and  $T$  time-points in latent class  $k_2$ , respectively (Viroli, 2011). The resulting multivariate covariance matrix satisfies three goals: capturing the heterogeneity between individuals, capturing the temporal correlation, and capturing the dependence between variables. Representing the temporal covariance matrix through a Kronecker product also gives opportunity to impose several restrictions, yielding a parsimonious model, to aid the model estimation and interpretation. Also, their model formulation allows the variance of the white noise error,  $\sigma^2^{(k_2)}$ , to vary across time-points. Anderlucci and Viroli (2015) illustrate their model by investigating cognitive functioning in an elderly sample. Individuals were assessed on three outcome variables at six time-points over a total time span of ten years. Their model specification for this application includes a time-point-specific trend-cycle,  $\beta_t^{(k_1)}$ , which describes the average response level on each outcome variable at that time-point within a latent class. The model by Anderlucci and Viroli (2015) places individuals thus on a common time-axis, where the time-points are seen as comparable across individuals.

The latent class VAR model proposed by Ernst et al. (2020) employs a conditional representation of the longitudinal data in their estimation procedure. This representation allows to model intensive longitudinal data with many time-points, and longitudinal data of unequal length between individuals. The model by Ernst et al. (2020) does not require individuals to be assessed on a common time-axis; rather this model lends itself to longitudinal studies where individuals of various ages sign up at different times and exit the study after completing different numbers of time-points, as is common in ecological momentary assessment. Ernst et al. (2020) applied their model to intensive longitudinal data of eight emotions assessed at over 70 time-points per individual. In this application, they describe six latent sub-populations that are homogeneous with regards to their emotion dynamics and their trend-cycles. The trend-cycles describe the average emotion levels at the three different measurement times during the day (i.e., morning, midday, and evening) within a latent class. The trend-cycle in this model specification represents therefore the cyclical trends of emotions over the course of a day.

While the two models reviewed above represent the VAR coefficients within a latent sub-population only through fixed coefficients, a mixture multilevel VAR model has been proposed that extends the latent class VAR model by Ernst et al. (2020). This is done by including random VAR coefficients to account for between-individual differences within the latent sub-populations (Ernst et al., Submitted-b). Ernst et al. (Submitted-b) employed their model, as shown in Table 5.1, to analyse ecological momentary assessment data that describes the emotions of a heterogeneous sample, which includes young and elderly adults. With their mixture multilevel VAR model, they aimed to stratify the heterogeneous sample into sub-populations that are homogeneous in trend-cycles and dynamics. Across model estimation with different numbers of components, two components consistently re-emerged: one component that consisted of almost entirely young participants, with only very few of the elder participants, and another component that consisted predominately of elderly adults. On top of providing a latent categorisation of individuals into components, the mixture multilevel VAR model also provided insight into the between-individual variation in trend-cycles and dynamics within the components, via the variance of the random effects.



Table 5.1: The rows of the table show the structure that the various model parts (as specified above) take in the models reviewed here. An index of  $i$  or  $t$ , or a superscript ( $k$ ) indicate that a parameter depends on the individual, the time-point, or on the latent categorisation, respectively. Empty cells indicate that a property is not present in this model. For simplicity of notation, symbols are consistently bold or not bold throughout this manuscript, whether a model is specified for multiple outcome variables can be inferred from Tables 5.2a and 5.2b.

	$\underline{\mu}_{i,t}^{(k)}$	$k_1$	$\underline{\xi}_{i,t}^{(k_2)}$	$k_2$	$\underline{\eta}_{i,t}^{(k_1)}$	$\tau^2$	$\tau^2$	$\rho_a^{(k_2)}$	$\bar{\rho}_{i,t}^{(k_2)}$	$g^{(k_2)}$	$\underline{\varepsilon}_{i,t}^{(k_2)}$	$\sigma^2$
Standard models:												
Random coefficient model (e.g., Skrondal & Rabe-Hesketh, 2008)	$\mathbf{x}_{i,t}\boldsymbol{\beta} + \mathbf{x}_{i,t}\bar{\boldsymbol{\beta}}_i, \quad \bar{\boldsymbol{\beta}}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$				$\eta_{i,t}$	$\tau^2$					$\varepsilon_{i,t}$	$\sigma^2$
Growth curve model (e.g., McArdle, 1988)	$\mathbf{x}_{i,t}\boldsymbol{\beta} + \mathbf{r}_{i,t}\boldsymbol{\gamma} + \mathbf{r}_{i,t}\bar{\boldsymbol{\gamma}}_i, \quad \bar{\boldsymbol{\gamma}}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$				$\eta_{i,t}$	$\tau^2$					$\varepsilon_{i,t}$	$\sigma^2$
VAR model (e.g., Lutkepohl, 2005)			$(\sum_{a=1}^p \underline{\xi}_{i,t-a}\rho_a^{(k_2)}) + \varepsilon_{i,t}$					$\rho_a$		$\beta_i$	$\varepsilon_{i,t}$	$\sigma^2$
Models included in the review:												
DeSarbo and Cron (1988)	$\mathbf{x}_i\boldsymbol{\beta}^{(k_1)}$	$k_1$			$\eta_{i,t}^{(k_1)}$	$\tau^2$					$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$
Leinik and DeSarbo (1999)	$\mathbf{x}_{i,t}\boldsymbol{\beta}^{(k_1)} + \mathbf{x}_{i,t}\bar{\boldsymbol{\beta}}_{i,t}^{(k_1)}, \quad \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)})$	$k_1$			$\eta_{i,t}^{(k_1)}$	$\tau^2$					$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$
Serang et al. (2016)	$\mathbf{r}_{i,t}\boldsymbol{\gamma}^{(k_1)} + \mathbf{r}_{i,t}\bar{\boldsymbol{\gamma}}_{i,t}^{(k_1)}, \quad \bar{\boldsymbol{\gamma}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)})$	$k_1$			$\eta_{i,t}^{(k_1)}$	$\tau^2$					$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$
Young and Hunter (2015)	$\mathbf{x}_i\boldsymbol{\beta} + \mathbf{r}_i\boldsymbol{\gamma}^{(k_1)} + \mathbf{r}_i\bar{\boldsymbol{\gamma}}_i^{(k_1)}, \quad \bar{\boldsymbol{\gamma}}_i^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)})$	$k_1$			$\eta_{i,t}^{(k_1)}$	$\tau^2$					$\varepsilon_{i,t}$	$\sigma^2$
Ng et al. (2006)	$\mathbf{r}_i\boldsymbol{\gamma}^{(k_1)} + \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)} + \bar{\boldsymbol{\gamma}}_{i,t}^{(k_1)}, \quad \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}_2^{(k_1)})$	$k_1$			$\eta_{i,t}^{(k_1)}$	$\tau^2$					$\varepsilon_{i,t}$	$\sigma^2$
Wang et al. (2012)	$\bar{\boldsymbol{\beta}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)}), \quad \bar{\boldsymbol{\gamma}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)})$	$k_1 = k_2$			$\eta_{i,t}^{(k_1)}$	$\tau^2$		$\rho_1^{(k_2)}$			$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$
Funatogawa and Funatogawa (2012)	$\mathbf{r}_{i,t}\boldsymbol{\gamma}^{(k_1)} + \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)}, \quad \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)})$	$k_1 = k_2$			$\eta_{i,t}^{(k_1)}$	$\tau^2$					$\varepsilon_{i,t}$	$\sigma^2$
Jongerling et al. (2015)	$\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}_i, \quad \bar{\boldsymbol{\beta}}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$				$\eta_{i,t}$	$\tau^2$		$\rho_1$	$\bar{\rho}_{i,t}$		$\varepsilon_{i,t}$	$\sigma^2 = \sigma^2 + \bar{\sigma}_i^2$
Bringmann et al. (2013)					$\eta_{i,t}$	$\tau^2$		$\rho_1$	$\bar{\rho}_{i,t}$	$\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}_i, \quad \bar{\boldsymbol{\beta}}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$	$\varepsilon_{i,t}$	$\sigma^2$
Corroto et al. (2012)	$\boldsymbol{\beta}_i^{(c_1)}$				$\eta_{i,t}$	$\tau^2$		$\rho_1$	$\bar{\rho}_{i,t}$		$\varepsilon_{i,t}$	$\sigma^2$
Driver and Voeckle (2018)	$\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}_i, \quad \bar{\boldsymbol{\beta}}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$				$\eta_{i,t}$	$\tau^2 = \tau^2 + \bar{\tau}_i^2$		$\rho_1$	$\bar{\rho}_{i,t}$		$\varepsilon_{i,t}$	$\sigma^2 = \sigma^2 + \bar{\sigma}_i^2$
Anderlucci and Viroli (2015)	$\mathbf{x}_i\boldsymbol{\beta}^{(k_1)}, \quad \bar{\boldsymbol{\beta}}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$	$k_1 = k_2$			$\eta_{i,t}$	$\tau^2$		$\rho_a^{(k_2)}$	$\bar{\rho}_{i,t}$		$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$
Ernst et al. (2020)	$\mathbf{r}_i\boldsymbol{\gamma}^{(k_1)}$	$k_1 = k_2$			$\eta_{i,t}$	$\tau^2$		$\rho_a^{(k_2)}$	$\bar{\rho}_{i,t}$		$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$
Ernst et al. (Submitted-b)	$\boldsymbol{\beta}^{(k_1)} + \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)}, \quad \bar{\boldsymbol{\beta}}_{i,t}^{(k_1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^{(k_1)})$	$k_1 = k_2$			$\eta_{i,t}$	$\tau^2$		$\rho_1^{(k_2)}$	$\bar{\rho}_{i,t}^{(k_2)}$		$\varepsilon_{i,t}^{(k_2)}$	$\sigma^2$

Table 5.2a: Data structures and special properties of some of the models discussed in this paper: Part I. Empty cells indicate either that the estimation procedure was not reported on, or that there are numerous options to estimate the model.

	Data structure <sup>a</sup>	Estimation procedure	Software used	Special properties
Standard models:				
Random coefficient model	Multiple time-points, multiple individuals, single outcome			
Growth curve model	Multiple time-points, multiple individuals, single outcome			$r_{it}$ can contain, for instance, a 1 for the intercept term and $t_i$
VAR model	Multiple time-points, single individual <sup>b</sup> , multiple outcomes			
Models included in the review (Part I):				
DeSarbo and Cron (1988)	Single time-point, multiple individuals, single outcome	EM algorithm		
Lenk and DeSarbo (1999)	Multiple time-points, multiple individuals, single outcome	MCMC	Mplus for SEM, OpenBUGS for Bayesian	We show here the linear approximation of the non-linear model through Taylor series expansion: $r_{it}$ contains partial derivatives of the growth function with respect to the fixed effects
Serang et al. (2016)	Multiple time-points, multiple individuals, single outcome	SEM and Bayesian estimation	R package mixtools (regmixEM.mixed function)	Growth curve model: $r_{it}$ can contain, for instance, a 1 for the intercept term, $t_i$ , and $t_i^2$
Young and Hunter (2015)	Multiple time-points, multiple individuals, single outcome	ECM algorithm		Includes random effects shared by all time-points measured on a given tissue of a given gene (i.e., $\tilde{\beta}_t^{(k_i)}$ )
Ng et al. (2006)	Multiple time-points (measured on multiple tissue samples), multiple individuals (i.e., genes), single outcome	ECM algorithm	EMMIXcontrasts R package (emmixwire function)	

<sup>a</sup>Outcome denotes the dependent variable(s).

<sup>b</sup>Alternative specifications of the VAR model are possible that encompass multiple individuals, these specifications either (1) assess only a single outcome, or (2) allow for no between-individual differences in the coefficients.

Table 5.2b: Data structures and special properties of some of the models discussed in this paper: Part II. Empty cells indicate either that the estimation procedure was not reported on, or that there are numerous options to estimate the model.

Models included in the review (Part II):			
Data structure <sup>a</sup>	Estimation procedure	Software used	Special properties
Wang et al. (2012)	ECM algorithm	EMMIXcontrasts R package (emmixwire function)	$r_{i,t} \gamma^{(k)}$ represents a Fourier series approximation to account for periodic trends
Funatogawa and Funatogawa (2012)	Newton-Raphson ridge method	Their SAS program is available upon request	Number of time-points may differ across individuals
Jongering et al. (2015)	Restricted maximum likelihood estimation (REML) and Bayesian estimation	lme4 R package for REML, winBUGS for Bayesian estimation	
Bringmann et al. (2013)	Pseudo-likelihood estimation for the covariance matrix between various outcome variables	lme4 R package	
Gorrostieta et al. (2012)	Empirical Bayes estimates (random effects) and conditional restricted maximum likelihood (fixed effects) [sic] <sup>b</sup>	proc mixed in SAS	
Driver and Voelkle (2018)	Hierarchical Bayesian modeling	cism R package (via stan)	Continuous time model
Anderlucci and Viroli (2015)	EM algorithm	Their R function is available upon request	Estimate the model-implied covariance matrix of the longitudinal data
Ernst et al. (2020)	EM algorithm	LCVAR R function <sup>c</sup>	$\mathbf{z}_i^{(k_s)} \sim \mathcal{N} \left( \mathbf{0}; \mathbf{\Sigma}_{within}^{(k_s)} \otimes \mathbf{\Sigma}_{between}^{(k_s)} \right)$
Ernst et al. (Submitted-b)	EM algorithm	Mplus	

<sup>a</sup> Outcome denotes the dependent variable(s).

<sup>b</sup> In a mixed linear model, fixed effect coefficients cannot be estimated with REML, because their term does not appear in the likelihood (see Section 1.6.2 in De Leeuw & Meijer, 2008). They probably mean a weighted least squares estimate (full information maximum likelihood estimate) of the fixed coefficients that uses the REML estimate of variance parameters as a weight matrix.

<sup>c</sup> Available at: <https://github.com/AnieBeel/LCVAR>.

### 5.3.6 Summary of the review

Above, we reviewed a number of models that account for between-individual differences through latent categorisations and/or random coefficients. In this review we illustrated through concrete examples how the various elements of our model framework are used in specific statistical models to answer empirical research questions. Our review illustrated that cross-sectional data models, trajectory models and dynamic models are vastly different in the data structures they can accommodate and the research questions they can answer — yet, these distinct models employ congenial approaches to describe between-individual differences in multisubject data, even across various disciplines. Originally, cross-sectional regression models were extended to include latent categorisations to identify discrete between-individual differences in the regression coefficients (DeSarbo & Cron, 1988). These models were extended further to include random effects, which describe continuous between-individual differences. The resulting models capture discrete and continuous between-individual differences and can accommodate longitudinal data (Lenk & DeSarbo, 1999). Later, trajectory models were extended in a similar way to become growth curve models (McArdle, 1988), which account for continuous between-individual differences, latent class growth analysis (Muthén & Asparouhov, 2008), which account for discrete between-individual differences, and to growth mixture models (Muthén & Asparouhov, 2008; Serang et al., 2016; Young & Hunter, 2015), which account for discrete and continuous between-individual differences. With the rise of intensive longitudinal data in various disciplines, dynamic models, specifically (V)AR models, became very popular. As a result, (V)AR models were extended to multilevel (V)AR models (Bringmann et al., 2013; Driver & Voelkle, 2018; Gorrostieta et al., 2012; Jongerling et al., 2015), which capture continuous between-individual differences, latent class (V)AR models (Anderlucci & Viroli, 2015; Ernst et al., 2020), which capture discrete between-individual differences, and mixture multilevel (V)AR models (Ernst et al., Submitted-b), which capture discrete and continuous between-individual differences. These (V)AR model extensions now shed additional insight into between-individual differences in dynamic processes. Our brief review demonstrated that despite differences in data structure and area of application, these models can all be expressed using the unifying model framework that we have proposed above. The expressions of these models in terms of our model framework are summarised in Table 5.1.

## 5.4 Extensions to the model framework

The model framework proposed in this paper is rather general, yet it can be extended even further. For instance, to include autoregressive conditional heteroskedasticity (ARCH) models, which model the white noise error variance,  $\sigma^2$ , explicitly (see e.g., Rast et al., 2020). Also, state-space extensions to our model framework are possible that describe the fixed coefficients' evolution across time (e.g., Gamerman & Migon, 1993). The focus of our model framework is on accounting for between-individual differences with random effects and latent categorisations. Therefore, we discuss in more detail below the extension to include latent sub-populations at different levels of the hierarchical nesting structure.

The model framework proposed in this paper can be seen as a special case of a multilevel regression mixture model that contains at the between-level two mixture models, containing several components each (Asparouhov & Muthén, 2008). In our model framework, time-points at level 1 are nested within individuals at level 2. An important consideration in multilevel mixture models is whether the latent sub-populations are assumed at the within-level and/or the between-level (Asparouhov & Muthén, 2008). When the mixture components are at the within-level (e.g., Muthén & Asparouhov, 2009), more advanced estimation procedures are needed, because random effects are included at the level of the mixture components. This

implies that subjects within a latent sub-population share the same random effects. This induces correlation within the latent sub-populations. This implies that the estimation requires a forward-backward EM algorithm where, as a consequence, the E-step has to be carried out conditionally on these components' random effects; the subjects are conditionally independently distributed given the component's random effects (Vermunt, 2003). Accounting for this dependency is not needed when latent sub-populations are assumed at the between-level. That is, when all the within-level observations that are nested in one between-level cluster (i.e., within a person), are assumed to arise from the very same latent sub-population (also denoted as stochastic constraint, see Asparouhov & Muthén, 2008). For between-level latent sub-populations, the question which information will be employed to predict latent sub-population membership, becomes more cumbersome. For instance, when only between-level variables can be employed to predict latent membership, researchers will be left with much fewer observations. In the modelling framework at hand, we use the within-level observations to determine latent sub-population membership at the between-level.

Our model framework could be extended further to allow for more complex nesting structures. Also, latent sub-populations could be assumed at the within-level, or at the between-level and the within-level (Muthén & Asparouhov, 2009). This could be done, for instance, by relaxing the stochastic constraint and allowing the time-points of an individual to arise from different latent sub-populations, as is done in hidden Markov modelling (Asparouhov et al., 2017). Further, in some situations, the individuals whose longitudinal data we observe, might be nested in groups. For instance, when students are sampled from different schools. In that case, models with latent classes or mixture components at different levels of the nesting structure might be desired (see, Muthén & Asparouhov, 2009; Vermunt, 2003, 2004; Vermunt & Magidson, 2005). For instance, the observed membership in a cluster at level 3 (i.e., school membership) might influence the latent categorisation at level 2 (i.e., at the individual-level, see e.g., Henry & Muthén, 2010).

## 5.5 Discussion

### 5.5.1 Concluding remarks on model specification

Above, we have specified a broad and general model framework for modelling between-individual differences in longitudinal data. We illustrated the usefulness of our model framework by demonstrating that a diverse set of longitudinal models can be expressed using our framework. This set includes some of the most prominent longitudinal models in education, psychology, sociology and biostatistics (i.e., random coefficient models, growth curve models, and VAR models). The model framework lends itself therefore to connect distinct multisubject longitudinal models to one another, to highlight their commonalities and differences.

Does the presentation of our broad model framework mean that we recommend to specify equally broad statistical models to analyse longitudinal data? Our answer to this question is a firm no — for reasons outlined now. In general, more parsimonious models that are tailored to the research question at hand, should be favoured. The more general and flexible a statistical model is, the harder it is to purposefully interpret the model and its results. For instance, the interpretation of any parameter included in a statistical model may change with any other parameter that is included into this statistical model. Similarly, when the statistical significance of a parameter in a model is tested, all other parameters that are included in the model serve as the null hypothesis (Snijders & Bosker, 2011a). Thus, parameters influence the interpretation and the statistical power of other parameters in the model. When specifying a statistical model, researchers should strive to specify the most parsimonious model that realises two simultaneous goals (1) describe the data sufficiently well, and (2) describe the

data in ways that are of substantive interest (Snijders & Bosker, 2011a). The former goal is related to model fit, the latter to answering research questions and specifying a sensible model based on the subject matter.

While there are theoretical advantages to a parsimonious model, such as the model being easier to understand and interpret, sometimes restrictions need to be imposed for a model to be identifiable and/or estimateable. For instance, covariates and intercept terms can be included either in  $f^{(k_1)}$  or in  $g^{(k_2)}$ , but not in both because such a model is not identifiable. A model including an intercept in  $f^{(k_1)}$  mean-centres the time series; this replaces the need for an intercept in  $g^{(k_2)}$ . For many longitudinal models, we prefer the inclusion of the intercept in  $f^{(k_1)}$  (i.e., estimating the within-person mean of the longitudinal data) over its inclusion in  $g^{(k_2)}$  (i.e., estimating a traditional (V)AR intercept), because (V)AR intercepts lack a meaningful interpretation (see e.g., Jongerling et al., 2015). Covariates included in  $f^{(k_1)}$  influence only one measurement, and these influences do not carry over to any future time-point. In contrast, the influence of covariates included in  $g^{(k_2)}$  carry over to the next time-points. When the influences of the covariates are assumed to carry over to the next time-point, covariates should be included in  $g^{(k_2)}$ , otherwise covariates should be included in  $f^{(k_1)}$ . To give another example, in a random coefficient model (as described in Table 5.1), random coefficients can only be applied to time-varying covariates; random coefficients for time-constant covariates are not identifiable. Similarly, fixed coefficients can be applied to time-varying and time-constant covariates as long as the model does not include a distinct fixed intercept coefficient for every person. In the latter case, fixed coefficients can be included only for time-varying covariates.

When estimating a model that is specified based on our model framework, imposing the constraint  $k_1 = k_2$  for all individuals will aid the estimation procedure. This constraint forces the latent categorisations for trend-cycles and for dynamics to be identical. Thereby the model assumes that individuals within a latent sub-population are homogeneous with regards to trend-cycles and dynamics. While it is possible to estimate two separate latent categorisations for the trend-cycles and the dynamics, meaning such a model is in principle identifiable, the estimation of these two distinct latent categorisations would be very difficult. As can be seen in Equation 5.1, the value of the dynamics are contingent on the value of the trend-cycles, and vice versa the value of the trend-cycles are contingent on the value of the dynamics. Thus in the estimation of a model that aims to identify latent categorisations in an exploratory fashion, the estimation of latent categorisations for the trend-cycles (i.e.,  $k_1$ ) and the estimation of latent categorisations for the dynamics (i.e.,  $k_2$ ) will be conditional on one another (for more details see Ernst et al., Submitted-b).

### 5.5.2 Concluding remarks for empirical researchers

The model framework we have introduced in this paper, was specified to aid statisticians and empirical researchers in navigating and comparing different multisubject longitudinal models. We hope this framework will equip empirical researchers with insight into the various building blocks that can be combined to specify multisubject longitudinal models. By combining these building blocks in a coordinated way, empirical researchers can specify models that are tailored to the data and the research question at hand. Depending on the building blocks specified, these models can account for between-individual differences to various extents, and describe between-individual differences in many of the multitudinous attributes that can be assessed with longitudinal data, such as growth, cyclical trends, and dynamics, to name just a few.

