MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 2, 362–392

A Gentle Introduction and Application of Feature-Based Clustering with Psychological Time Series

Jannis Kreienkamp[a], Maximilian Agostini[a], Rei Monden[b], Kai Epstude[a], Peter de Jonge[a,c], and Laura F. Bringmann[a,c]

[a]Department of Psychology, University of Groningen, Groningen, Netherlands; [b]Graduate School of Advanced Science and Engineering, Hiroshima University, Higashihiroshima, Japan; [c]Interdisciplinary Center Psychopathology and Emotion Regulation, Groningen, Netherlands

ABSTRACT

Psychological researchers and practitioners collect increasingly complex time series data aimed at identifying differences between the developments of participants or patients. Past research has proposed a number of dynamic measures that describe meaningful developmental patterns for psychological data (e.g., instability, inertia, linear trend). Yet, commonly used clustering approaches are often not able to include these meaningful measures (e.g., due to model assumptions). We propose feature-based time series clustering as a flexible, transparent, and well-grounded approach that clusters participants based on the dynamic measures directly using common clustering algorithms. We introduce the approach and illustrate the utility of the method with real-world empirical data that highlight common ESM challenges of multivariate conceptualizations, structural missingness, and non-stationary trends. We use the data to showcase the main steps of input selection, feature extraction, feature reduction, feature clustering, and cluster evaluation. We also provide practical algorithm overviews and readily available code for data preparation, analysis, and interpretation.

KEYWORDS

Time series analysis; feature-based clustering; intensive longitudinal data; ESM

Recent years have seen a striking increase in the number and variety of research studies that follow participants' everyday experiences and collect real-world psychological time series (e.g., Hamaker & Wichers, 2017). These intensive longitudinal datasets come with different sources of heterogeneity, where researchers have to consider differences across large numbers of participants, time points, and variables (e.g., Cattell, 1966; Wardenaar & de Jonge, 2013). However, despite its complexity, researchers are often interested in precisely this complexity and wish to understand how people differ in their developments across several variables (e.g., Ernst et al., 2021). Researchers and practitioners are, for example, asking: "Do the symptoms of different patients develop in contrasting ways?" (Monden et al., 2015) or "How do migrants differ in the development of their self-reported needs as they arrive in a new country?" (Kreienkamp et al., 2024). There is, thus, a clear need for analysis techniques that identify between-subject differences in developmental patterns for psychological data.

Recently, one promising way of identifying between-subject developmental patterns has been *time series clustering*—the idea of inductively grouping participants based on similarities of their time series (e.g., Ariens et al., 2020; also see den Teuling et al., 2021 for a review). This type of analysis essentially seeks to capture comparable within-person developments—such as whether a variable remains stable over time, consistently increases, or exhibits cyclical patterns—and then groups the persons based on these patterns (Liao, 2005). Time series clustering, thus, crucially depends on identifying meaningful summaries of the time series developments, which can be used to compare participants (Aghabozorgi et al., 2015).

Fortunately, past conceptual and empirical works in the experience sampling (ESM) literature have collected a number of meaningful aspects of

psychological time series.[1] Such aspects can be indicative of adaptive and maladaptive developments within the individual, can identify crucial transitions, or more generally are helpful in understanding a psychological time series. Important aspects might include concerns about whether a symptom consistently stays at a certain level without much variability or whether some emotions develop together. For the most important developmental aspects, researchers have assembled measures that capture these patterns. These summary statistics are often called "dynamic measures," "principles of change," or "dynamic features" of the psychological time series (Dejonckheere et al., 2019; Krone et al., 2018; Kuppens & Verduyn, 2017). Most research groups working on these time series features have proposed an overlapping number anywhere between four and twelve key features relevant to psychologists (Dejonckheere et al., 2019; Wang et al., 2006). Each of these time series features not only captures a distinct aspect of psychological time series but also holds conceptual value—inertia, for example, describes a resistance to change that can be indicative of psychological maladjustment (Kuppens et al., 2010) or a higher within-person variability can signal an erratic state (Myin-Germeys et al., 2018).

In this manuscript, we introduce *feature-based time series clustering*—a general clustering approach and framework where users utilize the dynamic features directly and can build upon readily accessible algorithms. The approach has been a common procedure in digital phenotyping (Loftus et al., 2022) and the broader machine learning literature (Maharaj et al., 2019). As such, the analysis has been applied to a variety of data, including analyses of astronomical, meteorological, and aviation pathways, biological and medical developments, as well as energy and finance patterns (Aghabozorgi et al., 2015). We argue that for psychological time series data, feature-based clustering offers a direct and flexible approach to use dynamic features, fewer strict assumptions than commonly used clustering approaches, beginner-friendly analysis methods, as well as a focus on meaningful psychological interpretability.

In the sections below, we aim to provide a practical introduction to the method. To do so, we illustrate the utility of the method with real-world ESM data. We use this data to discuss which psychological time series features are well-suited for a clustering approach, introduce the individual analysis steps, and provide practical guidance on common algorithms and analysis code. As such, we seek to reach readers who are familiar with ESM data but are relatively new to the realm of time series clustering. This article aims to introduce the important research decisions and focuses on approachable and accessible methods. As a result, the features and methods we choose to highlight might deviate from the state-of-the-art most advanced methods. For readers who would like to explore more specialized algorithms, we have prepared Supplemental Material C to embed the approach in the broader time series clustering literature and discuss alternative algorithms.

Why feature-based clustering for ESM

Let us briefly consider why you might consider feature-based time series clustering for your ESM data. We want to mention three key contributions that this introduction to the method seeks to highlight. The first contribution is that the feature-based approach aligns well with the growing literature on "dynamic features" of psychological ESM data (Dejonckheere et al., 2019; Krone et al., 2018; Kuppens & Verduyn, 2017). This match offers two main benefits, flexibility and interpretability. By relying directly on the time series features that are already well established for different psychological processes, users can mix and match the features that match their empirical assumptions and research questions. At the same time, the flexibility does not reduce the interpretability of the results because everything that goes into the clustering process has a clear conceptual meaning (for more details on this see the 'Feature Extraction' step below).

The second contribution is that the direct use of time series features avoids many of the challenges that psychological time series data face today. Researchers frequently seek to address important questions about multidimensional, erratic, and context-specific phenomena (Hamaker & Wichers, 2017; Helmich et al., 2020; Kivelä et al., 2022). These types of data often include issues of non-equidistant, structurally missing, or non-stationary data (also see Appendix A for an expanded discussion of the current challenges within ESM data). Time series features, however, commonly do not require complete and equidistant time series, and the approach actively encourages the inclusion of

[1]In psychology, intensive longitudinal data collection methods are often referred to as experience sampling method (ESM), ecological momentary assessment (EMA), or ambulatory assessment (AA) studies. Although the terms come from different conceptual backgrounds, they share a focus on collecting data over an extended period of time to capture people's behaviors and experiences as they vary over time and in response to different situations and events. In this article, we will use the experience sampling (ESM) term as it has the strongest footing within the clustering literature.

non-stationary and non-linear trends (e.g., Aghabozorgi et al., 2015).

As such, the feature-based approach extends the ESM clustering approach most commonly used today. Most ESM research today clusters participants based on person-specific model parameters—notably intercepts and slopes from vector autoregression models (VAR; e.g., Ariens et al., 2020; Bulteel et al., 2016; Stefanovic et al., 2022). The utility of these model parameters is notably high because they often fit well within a research process where model parameters are already estimated for theory testing. However, the model parameters crucially depend on the features included in the model and the assumptions of the model.[2] Feature-based time series clustering considers model parameters to be one type of time series feature but also allows users to include other time series features without needing to develop or fit another (non)parametric model. Users might thus use VAR parameters alongside piecewise polynomials, the *Mean of the Squared Successive Differences* (*MSSD*), or simply the participants' variance estimates (see the *Feature Reduction* section below).

The third major contribution is that the feature-based time series approach offers a generalized structure for clustering ESM data. The generalized decision structure of feature-based clustering provides a methodological framework that also applies to most clustering approaches already used within the ESM literature, with the added benefit of having practical embeddedness from other disciplines (Liao, 2005). As part of the illustration, we separate the analysis process into feature extraction, feature reduction, feature clustering, and cluster evaluation (Räsänen & Kolehmainen, 2009; Wang et al., 2006). This approach crucially gives (novice) users guidance on where researcher decisions need to be made, and we aim to provide information on how to approach these decisions. Importantly, the generalized structure can also capture the existing model parameter-based analyses (e.g., Ernst et al., 2021) as well as other approaches that rely on features more directly (e.g., van Genugten et al., 2022).

Data used for illustration

To illustrate the functioning and utility of feature-based time series clustering with psychological ESM data, we introduce the clustering process using a recent set of studies that collected data on migration experiences. We chose this particular set of data because they exemplify the key contributions and challenges that we seek to address with the feature-based approach. In particular, the use of ESM for migrants in interactions with the majority society has a long-standing tradition of dynamic theorizing (e.g., Berry, 1986) with multivariate conceptualizations (e.g., Kreienkamp et al., 2024), a focus on non-stationary trajectories (Kim, 2017), distinguishing adaptive from maladaptive clusters (Choi et al., 2009), and event-based missingness (e.g., Keil et al., 2020; Wardenaar & de Jonge, 2013).

Matching these requirements, the data set we use consists of three studies that followed migrants who had recently arrived in the Netherlands in their daily interactions with members of the Dutch majority group (for the data set see Kreienkamp et al., 2022). After a general migration-focused pre-questionnaire, participants were invited twice per day to report on their (potential) interactions with majority group members for at least 30 days. The short ESM surveys were sent out at around lunch (12 pm) and dinner time (7 pm). After the 30-day study period, participants filled in a post-questionnaire that mirrored the pre-questionnaire. Participants received either monetary compensation or partial course credits based on the number of surveys they completed.

The original studies included 207 participants ($N_{S1} = 23$, $N_{S2} = 113$, $N_{S3} = 71$) with a total of 10,297 ESM measurements. Each of the studies focused on newly arrived first-generation migrants, and each study included a number of idiosyncratic variables relevant for the broader research collective. For our empirical example, we focus on the variables that were collected during ESM surveys and were available in all three studies. Variable selection and preparation are described as part of the illustration below, but for additional methodological details about the study setup, see Kreienkamp et al. (2022). Each study was approved by the ethics board of the university of origin and all participants gave their informed consent.

Analysis steps and application

To introduce and illustrate the feature-based clustering analysis, we will follow the conceptual steps of the procedure in sequential order and discuss key issues for each step. To do so, we use a framework that structures feature-based clustering into four main steps (Räsänen & Kolehmainen, 2009; Wang et al.,

[2]Importantly, recent efforts to address the shortcomings of model parameters for time series clustering have made notable progress by either relaxing specific assumptions (e.g., Chow et al., 2011; den Teuling et al., 2021; Molenaar et al., 2009; Ou et al., 2023; Voelkle & Oud, 2013) or include additional time series features (e.g., see Gates et al., 2017; Krone et al., 2018).

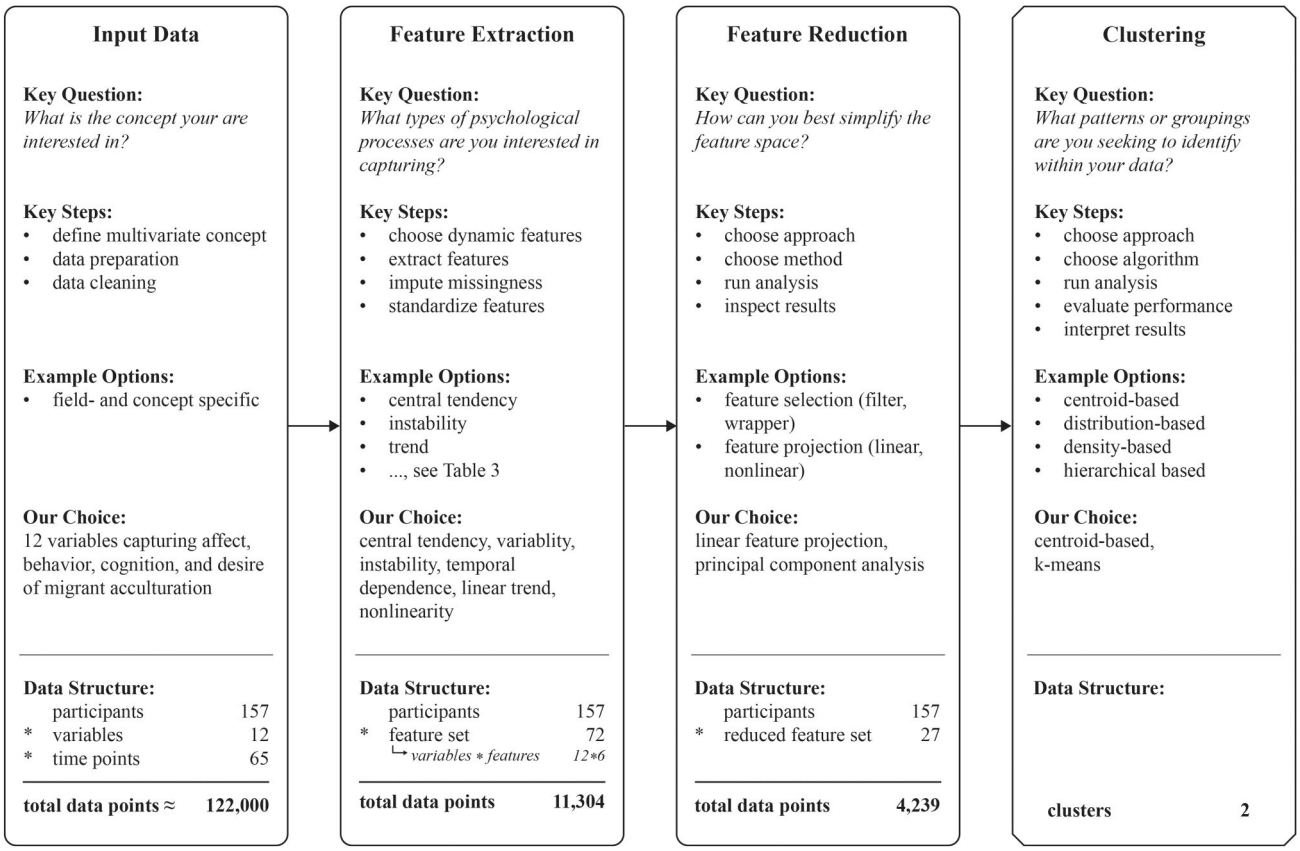

Figure 1. Flowchart feature-based time series clustering in psychology.

2006). (1) The selection and preparation of the input variables, (2) the extraction of the time series features that describe the time series, (3) an optional feature reduction step if there are too many data points for the clustering algorithms, and (4) the actual clustering of the time series features, as well as the evaluation and interpretation of the clusters. While this is a useful conceptual separation of procedural elements, it is important to note that these steps are a general outline and the specific details of the analysis will depend on the nature of the data and the research question being addressed. We, thus, mainly wish to highlight the conceptual nature of these steps to introduce the major elements of the analysis. We also provide a conceptual overview that can be used in conjunction with this section in Figure 1.

Input variables

Time series clustering starts with the selection and preparation of the variables of interest. While the selection will necessarily be field- and concept-specific, there are a few conceptual and methodological issues that should be considered. Conceptually, the included variables should adequately capture the concept of interest and should be meaningful to the understanding of the time series. One of the advantages of feature-based clustering is that it is inherently adept at accommodating multivariate concepts, a common aim in ESM research. For example, there are calls that emotion dynamics should be assessed with a repertoire of positive and negative emotions (e.g., Dejonckheere et al., 2019), many health developments are captured within biopsychosocial domains (e.g., Suls & Rothman, 2004), and migration experiences are thought to encapsulate affect, behavior, cognition, and desire measurements (e.g., Kreienkamp et al., 2024). At the same time, however, the added number of variables can become a methodological concern. Not only can redundant and irrelevant variables diminish the quality of the analyses, but with intensive longitudinal data the number of data points compounds across participants, measurement occasions, and variables so that additional variables can make many of the following steps substantially more difficult (also see the 'Data Structure' in Figure 1).

For our illustration, we include 12 variables that were measured as part of the ESM surveys in all three studies and captured information about the participant's interactions, as well as the cognitive-, emotional-, and motivational self in relationship with the majority group (see Table 1 for an overview). We chose these

Table 1. Variable selection.

Variable	Question	Aspect	ESM	Contact specific: Interaction only	Contact specific: unspecific
Int: Accidental	The interaction with -NAME- was accidental.	Cognition	✓	✓	
Int: Cooperative	The interaction with -NAME- was cooperative.	Cognition	✓	✓	
Int: Meaningful	Overall, the interaction with -NAME- was: Superficial—Meaningful	Cognition	✓	✓	
Int: Need Fulfillment	During your interaction with -NAME- your goal (-GOAL-) was fulfilled.	Needs	✓	✓	
Int: Need fulfillment partner	-NAME- helped fulfill your goal (-GOAL-)	Needs	✓	✓	
Int: Partner attitude	At the moment, how favorably do you feel toward -NAME-	Cognition	✓	✓	
Int: Quality	Overall, the interaction with -NAME- was: Unpleasant—Plesant	Cognition	✓	✓	
Int: Representative	The interaction with -NAME- was representative of the Dutch.	Cognition	✓	✓	
Int: Voluntary	The interaction with -NAME- was voluntary.	Cognition	✓	✓	
Need fulfillment	During this -morning/afternoon- your goal (-GOAL-) was fulfilled.	Needs	✓		✓
Outgroup attitude	At the moment, how favorably do you feel toward the Dutch.	Cognition	✓		✓
Well-being	How do you feel right now? very sad—very happy	Emotion	✓		✓

Note: "Int" = interaction.
All items used a continuous slider and were rescaled to a range of 0–100.

aspects in particular because (1) the interaction-specific information exemplified the structural missingness issue of modern ESM data (see Appendix A for more detail) and (2) the motivational, emotional, and cognitive experience offered a diverse conceptualization of migration experience (beyond behavioral measurements) that is becoming more common in the literature (Kreienkamp et al., 2024). The breadth of the included variables also showcases the utility of the method for a growing body of literature that considers heterogeneous and complex concepts. As a result, the number of included variables is also on the higher end for psychological concepts and additionally allows us to showcase the efficiency benefits of the method and offers a reasonable use case for the feature reduction step.

Once the important variables have been selected, the data needs to be prepared for the analysis steps. Importantly, this not only means validating and cleaning the data (e.g., re-coding, combining scale items) but also making the time series comparable. Making time frames and response scales comparable between participants, for example, includes choosing a time frame that is common to most participants ('data preparation' and 'data cleaning' in Figure 1; also see Liao, 2005).

In our illustration data set, the studies differed substantially in the maximum length of participation ($\max(t_{S1}) = 63$, $\max(t_{S2}) = 69$, $\max(t_{S3}) = 155$). This was likely due to the option to continue participating without compensation in the latter study. To make the three studies comparable in participation and time frames, we iteratively removed all measurement occasions and participants that had more than 45% missingness (which was in line with the general recommendation for data that might still need to rely on imputations for later model testing; see Madley-Dowd et al., 2019).[3] This procedure led to a final sample of 157 participants, who jointly produced 8,132 survey responses. Importantly, both the participant response patterns and the time frame were now substantially more comparable (number of measurement occasions per person: t_{S1} : min = 40, max = 61, mean = 57.33, sd = 4.69; t_{S2} : min = 33, max = 60, mean = 49.05, sd = 6.73; t_{S3} : min = 36, max = 65, mean = 54.20, sd = 7.04). It is important to consider that some time series features may be less reliable when the number of measurement occasions per person is low (e.g., below 30

[3]Please note that for cases where the clustering is the main analysis, this high missingness threshold may be too conservative. As part of our validation analyses in Appendix B we compare the model presented here with varying levels ofmissing data allowed.

Table 2. Correlation table and descriptive statistics.

	Int: Accidental	Int: Voluntary	Int: Cooperative	Int: Representative	Int: Meaningful	Int: Quality	Int: Need Fulfill.	Int: Need Fulfill. Partner	Attitude Partner	Daytime Core Need	Outgroup Attitude	Well-being
Correlations												
Int: Accidental		−0.16	−0.18	0.05	−0.08	−0.08	−0.33**	−0.22*	−0.03	−0.25*	−0.02	0.18
Int: Voluntary	−0.14***		0.60***	−0.07	0.07	0.39***	0.15	0.14	0.38***	0.08	0.21*	−0.06
Int: Cooperative	−0.14***	0.28***		0.15	0.36***	0.65***	0.41***	0.52***	0.44***	0.24*	0.24*	−0.08
Int: Representative	0.00	0.07**	0.12***		0.09	0.12	−0.08	−0.06	0.12	−0.01	0.41***	−0.05
Int: Meaningful	−0.19***	0.21***	0.29***	0.01		0.64***	0.07	0.17	0.42***	0.07	−0.04	−0.05
Int: Quality	−0.09***	0.32***	0.39***	0.06**	0.44***		0.40***	0.31**	0.63***	0.21	0.21*	0.10
Int: Need Fulfillment	−0.08***	0.18***	0.26***	0.10***	0.17***	0.32***		0.64***	0.08	0.64***	0.14	0.33***
Int: Need Fulfillment Partner	−0.11***	0.21***	0.32***	0.08***	0.20***	0.32***	0.52***		0.10	0.51***	0.11	0.08
Attitude Partner	−0.05	0.30***	0.30***	0.03	0.41***	0.58***	0.23***	0.26***		−0.13	0.55***	0.11
Daytime Need Fulfillment	−0.06*	0.11***	0.17***	0.02	0.16***	0.17***	0.15***	0.14***	0.10***		0.05	0.15
Outgroup Attitude	−0.03	0.14***	0.16***	0.15***	0.20***	0.31***	0.19***	0.22***	0.37***	0.09***		0.25**
Well-being	−0.05*	0.16***	0.17***	−0.03	0.22***	0.32***	0.15***	0.14***	0.26***	0.20***	0.24***	
Descriptives												
Grand Mean	39.10	80.08	79.55	64.65	61.16	79.85	85.42	78.52	80.59	76.48	66.84	74.82
Between SD	31.14	20.61	18.41	21.12	24.62	17.05	16.01	21.53	16.33	21.63	18.54	15.97
Within SD	28.72	19.27	17.43	19.92	22.32	16.37	18.63	20.02	15.81	22.26	9.45	12.86
ICC(1)	0.21	0.29	0.27	0.35	0.31	0.25	0.18	0.26	0.25	0.20	0.77	0.52
ICC(2)	0.90	0.93	0.93	0.89	0.94	0.92	0.91	0.92	0.91	0.92	0.99	0.98

Note: "Int." = outgroup interaction, "ICC" = intraclass correlation coefficient, and "SD" = standard deviation.
Upper triangle: Between-person correlations.
Lower triangle: Within-person correlations.
***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

measurements per person), and this should be taken into account when conducting similar analyses. Full methodological details are available in Online Supplemental Material A, but basic item information, descriptives, and correlations of items are also available in Table 2.

Feature extraction

Armed with a relevant selection of key variables, the main aim of the feature extraction is to describe the most important and meaningful aspects of a time series. In its most general approach, feature extraction can include any numeric summary of the time series (e.g., Maharaj et al., 2019). Given this flexibility, a staggering variety of time series features have been proposed across different disciplines. For example, Wang et al. (2006) proposed 9 time series features (also see Fulcher et al., 2013), Adya et al. (2001) collected 28 features relevant for forecasting, and a commonly used software package for feature extraction 'tsfresh' allows users to extract a total of 794 features of a time series (Christ et al., 2018).

However, not all time series features might be relevant to psychological time series or any particular research question. For example, a psychologist interested in well-being might not necessarily be interested in the exact time point after which 50% of the summed well-being values lie (i.e., relative mass quantile index) or how much different sine wave patterns within the well-being data correlate with one another (i.e., cross power spectral density). Instead, we advocate that we look at time series features that have a strong backing within the ESM literature and offer meaningful interpretability.

Fortunately, past conceptual and empirical efforts offer valuable discussions of common time series features in psychological research. To understand emotion dynamics, Kuppens and Verduyn (2017) originally proposed four dynamic features: (1) within-person variability, (2) co-variance or intraclass coefficient (ICC), (3) inertia or autocorrelation, and (4) cross-lagged correlations. These features were then extended by Krone et al. (2018), adding (5) innovation variance, and (6) mean intensity. Krone et al. (2018) even built a parametric model to tentatively cluster study participants. From a slightly different perspective Dejonckheere et al. (2019) later added three additional features for psychological time series: (7) instability (8) interdependence (i.e., network density),

and (9) diversity (i.e., Gini coefficient; also see Wendt et al., 2020).[4]

Some of the time series features found in the psychological literature are not necessarily well-suited to summarize time series for feature-based clustering and some key conceptual features are not well represented in the dynamic measures literature. In particular, covariances and cross-lagged correlations often produce a large number of parameters and can lead to overfitting (e.g., Bulteel et al., 2018; Ernst et al., 2021; Lafit et al., 2022), also see our discussion of multivariate features). Other dynamic features, such as network density parameters, used to summarize variable interdependence, may not always be meaningful for psychological data (Bringmann et al., 2019). At the same time, the dynamic features commonly proposed for psychological time series often do not capture linear and nonlinear trends, as they are frequently developed for stationary Vector Autoregressive models (e.g., Krone et al., 2018).

Thus, while the final selection of time series features should always be driven by the research questions and field-specific conventions, for our illustration we chose six time series features that relate to common psychological research questions and recent works within the field: (1) central tendency, (2) variability, (3) instability, (4) temporal dependence, (5) linear trend, and (6) nonlinearity. An exemplary overview of available time series features, their substantive interpretations, and mathematical operationalizations is available in Table 3, including the features we chose here. For each of the six time series features, we selected a mathematical representation that was appropriate for our type of data. We provide a brief introduction to each feature below. Beyond the operationalizations we chose for our case study, we collected the R functions we created for the analyses as an R package that automatically extracts and prepares a large selection of the time series feature operationalizations presented in Table 3. All functions are available as part of the package GitHub repository (see the featureExtractor() function; Kreienkamp et al., 2023d) and are annotated as part of our tutorial-style illustration (see Supplemental Material A).

Central tendency. The central tendency refers to the statistical measures that represent the "typical" or

Table 3. Examples of features for psychological time series.

	Substantive interpretation	Example operationalizations	Further reading
Central tendency	Average level of the experience across the entire measurement period.	• Mean • Median • Mode	Bringmann and Eronen (2018) and Weisberg (1992)
Variability	Describes the average deviation from the central tendency across the entire measurement period.	• Standard deviation • Variation coefficient • Median absolute deviation	Helmich et al. (2020) and van de Leemput et al. (2014)
(In)stability	Describes the average change between two consecutive measurements of the experience.	• Mean squared successive differences • Mean absolute change • Ix instability index	Kivelä et al. (2022), Wichers et al. (2019), and Wang et al. (2012)
Temporal dependence	Describes the extent to which current experiences or measurements are influenced by previous states or measurements. This includes resistance to change (i.e., carries over to the next measurement) and periodic or seasonal returns (e.g., self-predictive on a daily or weekly basis).	• Autocorrelation (e.g., lag–1) • Fourier coefficients • Continuous wavelet transform	Kuppens et al. (2010) and Walls et al. (2006)
Linear trend	Describes upwards or downwards linear trend of the experience reports.	• OLS regression slope • Avg. piecewise linear reg. slope	Gottman et al. (1969) and Oravecz et al. (2016)
Nonlinearity	Describes the nonlinear structure of the time series. This includes measures that indicate the deviation from the a linear trend as well as nonlinear model parameters.	• GAM spline edf • Bicoherence metrics • Langevin polinomial coefficient	Caro-Martín et al. (2018) and Bringmann et al. (2017)
Multivariate relations	Explores interactions and dependencies between multiple time series variables, capturing the complex behavior of psychological constructs as they evolve together over time (see discussion).	• Co-variance • (Graphical) VAR parameters • Network density	Epskamp et al. (2018) and Lacasa et al. (2015)

Note: The presented features and operationalizations are neither exhaustive nor necessary for feature-based clustering.

[4]It should be noted that also within the psychological literature, alternative summaries have been proposed that, for example, include measurement distribution, nonlinear developments, or categorical states. As an example, Kiwuwa-Muyingo et al. (2011) proposed to extract clinicallymeaningful states for medical adherence data and suggests these states as meaningful time series features.

"average" of a set of data. The most common measures of central tendency are the mean, median, and mode (Weisberg, 1992). As a familiar statistic from probability theory, the central tendency sits at the heart of many fundamental questions about psychological time series. Researchers might, for example, be interested in whether "Over a one-month period, are some people happier than others?"

For the central tendency feature of our illustration, we chose the median (M), which effectively addresses potential complications arising from non-normally distributed responses or outliers within time series datasets (Weisberg, 1992). To compute the median, it is imperative to differentiate between two types of time series representation for a given variable j related to participant i: the chronological series and the ordered series. The chronological time series, denoted by X_{ij}, encapsulates the sequence of observations $\{x_{ij1}, x_{ij2}, ..., x_{ijT}\}$ for variable j concerning participant i, organized by their temporal occurrence. Here, x_{ijt} signifies a specific observation at time t within this sequence. In contrast, the ordered time series, represented as $\mathbf{X}_{ij}$, is derived from X_{ij} by sorting the observations in ascending order of magnitude. This ordered set is expressed as $\{\mathbf{x}_{ij1}, \mathbf{x}_{ij2}, ..., \mathbf{x}_{ijn}\}$, with each $\mathbf{x}_{ijk}$ corresponding to the k-th element in the reordered series $\mathbf{X}_{ij}$.

The median $M(\mathbf{X}_{ij})$ is then the value located precisely at the center of the ordered time series $\mathbf{X}_{ij}$. Depending on whether the total number of observations (T) is odd or even, the median is either the middle k-th element if T is odd, or the average of the two middle values if T is even:

$$M(\mathbf{X}_{ij}) = \begin{cases} \mathbf{x}_{ij}\left(\frac{T+1}{2}\right) & \text{if } T \text{ is odd} \\ \frac{\mathbf{x}_{ij}\left(\frac{T}{2}\right) + \mathbf{x}_{ij}\left(\frac{T}{2}+1\right)}{2} & \text{if } T \text{ is even} \end{cases} \quad (1)$$

This approach ensures that the median is a reliable indicator of central tendency in time series analysis, unaffected by data distribution asymmetries or the presence of outliers.

Variability. Variability captures the degree to which a set of data differs from the central tendency and is sometimes also referred to as the dispersion or spread of the data (Weisberg, 1992). Common measurements of variability are the variance or standard deviation as well as their robust counterparts. In time series analyses, variability is conceptually important because information about the distribution and diversity of data has been found to be indicative of worse psychological states (Helmich et al., 2021; Myin-Germeys et al., 2018). Person-level differences in ESM measurements have, for example, been associated with higher levels of psychopathological recurrences among patients with depression (Timm et al., 2017). As such, psychological researchers and practitioners are often empirically interested in between-person differences in variability. Researchers on polarization and radicalization might, for example, ask: "Are people settled in their attitudes toward migrants or do they vary across the measurement period?"

For our illustration data, we chose the *Median Absolute Deviation* (*MAD*) to gauge the variability within our time series data. This choice is motivated by the robustness of MAD, particularly its resilience to the effects of non-normal distributions and outliers, which can significantly skew traditional variability measures such as the standard deviation (Weisberg, 1992). For a given variable j and participant i, the MAD is calculated by first determining the median (M) of the ordered time series $\mathbf{X}_{ij}$ as described in Equation (1). We then compute the absolute deviations of each observation in the time series X_{ij} from this median value. Specifically, for each time point t, we calculate the absolute difference between x_{ijt} and the median of the series $M(\mathbf{X}_{ij})$. The MAD is then the median of these absolute deviations:

$$MAD(X_{ij}) = M(\left|x_{ijt} - M(\mathbf{X}_{ij})\right|) \quad (2a)$$

$$= M(\{|x_{ij1} - M(\mathbf{X}_{ij})|, |x_{ij2} - M(\mathbf{X}_{ij})|, ..., |x_{ijn} - M(\mathbf{X}_{ij})|\}) \quad (2b)$$

The calculation of MAD focuses on the magnitudes of deviations, ensuring that it provides a robust measure of dispersion that reflects the inherent variability in the time series data.

Instability. Instability captures the average change between two consecutive measurements (Ebner-Priemer et al., 2009; Jahng et al., 2008). While instability is conceptually related to the variability feature, variability does not take into account temporal dependency, whereas instability looks at the 'jumpy-ness' of the data over time. In other words, variability reflects the range or diversity of values in the un-ordered time series data, while instability reflects the fluctuation or inconsistency in a time series data over time (Houben et al., 2015; Koval et al., 2013; Trull et al., 2008). For example, if a person has rapid and extreme mood changes, their mood is highly unstable, while if a person's mood responses span a wide range over the entire study period, their mood is highly variable (note that this does not need to be rapidly changing or instable, e.g.,

when there is linear increase over time; also see Jahng et al., 2008). Within psychological time series, instability measurements have especially been important in the research of borderline personality disorder (Trull et al., 2008) and suicidality (Kivelä et al., 2022), but also in understanding early warning signals more generally (Wichers et al., 2019). Conceptually, the instability feature, thus, relates to a broad range of research questions, including: "What is the nature of the identification changes in those who start working in a new country?" or "Do strong daily fluctuations in self-esteem reflect the process of identity formation in adolescents?"

For our data we chose the *mean absolute change* (*MAC*; e.g., Ebner-Priemer et al., 2009; Barandas et al., 2020), which looks at the average absolute difference of two consecutive measurements x at time points t and $t-1$, for each time series X of participant i and variable j.

$$MAC(X_{ij}) = \frac{1}{n-1} \sum_{t=2,\ldots,t} |x_t - x_{t-1}| \tag{3}$$

Another common measurement of instability is the *Mean of the Squared Successive Differences* (*MSSD*), which is often preferred where differences in magnitude are more important than the frequency of those changes, for example, when big shifts in time series are considered more impactful or when outliers are meaningful and need to be taken into account (Bos et al., 2019; Chatfield, 2003). For psychological ESM data, some research suggests that amplitude and frequency could predict different health outcomes and can be investigated jointly (Jahng et al., 2008; Wang & Grimm, 2012).

Temporal dependence. Univariate temporal dependence in time series data refers to the degree to which a time series is influenced by its past values, exhibiting patterns of behavior that may be regular over different time scales (D'Mello & Gruber, 2021). In the context of psychological time series, an important aspect of temporal dependence is *inertia*—how much a measurement carries over to its next measurement (Kuppens et al., 2010; Suls et al., 1998). If inertia is high, a development tends to stay in a certain state. Because high inertia is resistant to change, in emotion dynamics, high inertia of negative affect has been found to be indicative of under-reactive systems and to be characteristic of psychological maladjustment (Kuppens et al., 2010). In a similar vein, high inertia in negative affect at baseline was predictive of the initial onset of depression (Kuppens et al., 2012). Conceptually, inertia is more broadly connected to research questions such as: "Do patients stay in a negative mood for several measurements?" or "Do migrants stay with their language practice for several days at a time?" Note that we described univariate temporal dependence here, where the focus is on the relationship of a variable with its own past values. Cross-lagged effects extend this concept by examining how past values of one variable influence another.

For our illustration case, we chose the commonly used (univariate) autocorrelation or autoregression with a lag-1 to capture the inertia. High autocorrelation values can indicate high levels of inertia, while low autocorrelation values may indicate a more unpredictable or volatile time series (Dejonckheere et al., 2019). The lag–1 autocorrelation $r_{ij,1}$ looks at the average correlation between a measurement x and the preceding measurement x_{t-1} for the time series X of participant i and variable j with n measurements.

$$r_{ij,1} = \frac{\sum_{t=2}^{n}(x_{ijt} - \bar{x}_{ij})(x_{ij,t-1} - \bar{x}_{ij})}{\sum_{t=1}^{n}(x_{ijt} - \bar{x}_{ij})^2} \tag{4}$$

Where $\bar{x}_{ij}$ is the mean of the time series x_{ij}, calculated as:

$$\bar{x}_{ij} = \frac{1}{n}\sum_{t=1}^{n} x_{ijt} \tag{5}$$

While inertia captures the simplest case of temporal dynamics, lag-1, we acknowledge that temporal dependence in psychological time series may also exhibit more complex relationships, including higher lagged auto correlations or cyclical relationships (fourier coefficients, or continuous wavelet transforms are often used to capture such relationships).

Linear trend. In non-stationary time series, a linear trend can be observed when there is a consistent increase or decrease in the data over time (Nyblom, 1986). For psychological time series, researchers have, for example, pointed out the importance of linear trends in interpersonal communication (Vasileiadou & Vliegenthart, 2014), and emotion dynamics (Oravecz et al., 2016). Theoretically, linear trends are often considered the simplest way to assess whether a psychological theory of change is appropriate (Gottman et al., 1969). In empirical practice, linear trends are, thus, commonly exemplified by research questions such as "Do patient symptoms improve consistently?" or "Does worker productivity decline continuously?"

For the variables in our illustration data set, we chose an overall linear regression slope to capture the linear trend. The regression slope b_{ij} provides the average change from one time point t to the next across all measurements x of a time series X of

participant i and variable j. The specific form of the OLS slope formula we provide below calculates b_{ij} as the sum across all time points of the product of the deviation of time t from its mean $\bar{t}$ and the deviation of x_{ij} from its mean $\bar{x}_{ij}$ at each time point, divided by the sum across all time points of the square of the deviation of time from its mean ($\sum (t-\bar{t})^2$). Intuitively, the formula captures the rate of change of variable x_{ij} with respect to time. This slope will indicate how the variable x_{ij} changes over time, controlling for its mean value and the mean of time. If the slope is positive, x_{ij} increases over time; if it is negative, x_{ij} decreases over time.

$$b_{ij} = \frac{\sum (t-\bar{t})(x_{ijt}-\bar{x}_{ij})}{\sum (t-\bar{t})^2} \tag{6}$$

Nonlinearity. Changes in psychology are not always linear; instead, nonlinearity is a common feature of psychological time series (Hayes et al., 2007). As an example, episodic disorders, such as depression, are often best described as non-linear systems (Hosenfeld et al., 2015). Similarly, patients recovering from depression showed sudden changes in the improvement of depression (Helmich et al., 2020). But also substance abuse (Boker & Graham, 1998) or attitude changes rarely develop linearly (van der Maas et al., 2003). Conceptually, researchers might have research questions about the type of the development: "Is the development of well-being a nonlinear process?" as well as the shape and structure of the development: "How many spikes in well-being did a migrant experience?"

We summarized the nonlinear trend with the *estimated degrees of freedom* of an empty GAM spline model. The *edf* summarizes the *wiggliness* of a spline trend line (Bringmann et al., 2017; Wood, 2017). The degrees of freedom of a spline model are determined primarily by the number of knots and the order of the spline. For instance, a cubic spline with k knots has $k+3$ degrees of freedom (Castro-Alvarez et al., 2024; Faraway, 2016; Haslbeck et al., 2021). However, in a penalized spline framework, which is commonly used for GAMs, the effective degrees of freedom can be less than $k+3$. This is because the model employs a smoothing parameter to control the tradeoff between the complexity (flexibility) of the model and its fit to the data, thereby penalizing overly complex models and potentially reducing the effective degrees of freedom (Marx & Eilers, 1998). Intuitively, then an edf of 1 would be equivalent to a linear relationship (i.e., one linear slope parameter), whereas a higher edf (particularly an edf > 2) is indicative of a non-linear trend. The estimated degrees of freedom are commonly based on a concept called 'effective degrees of freedom' and can be represented as the trace $tr(\ )$, (i.e., the sum of the diagonal elements) of the smoother matrix S, a symmetric matrix that maps from the raw data to the smooth estimates (Wood, 2017).

$$edf = tr(S) \tag{7}$$

Additional considerations. Beyond our main features of interest, we also extracted the participant's number of completed ESM measurements to ensure that the clusters are comparable in that regard (i.e., to exclude spurious explanations for the cluster assignments). After the extraction of the features, we found that about 1.40% of the extracted features are missing across the 72 features per participant. This could happen, for example, if participants do not have two subsequent measurements with outgroup interactions, so that an autocorrelation with lag-1 cannot be calculated for the contact-specific variables. The small number of missing values indicates that the feature-based approach indeed largely avoids the structural missingness issue. However, even the few missing values can be an issue for some feature reduction or feature clustering algorithms. We, thus, impute missing feature values *via* predictive mean matching (PMM) with the MICE package in R, employing a single imputation and specifying a maximum of 50 iterations and a fixed seed for convergence and reproducibility (Buuren & Groothuis-Oudshoorn, 2011). We chose PMM for its ability to preserve the original data distribution without assuming normality and robustly handling multiple data types (Van Buuren et al., 2006). Note again that with this procedure we only need to impute an extremely small number of missing values, as most feature calculations can use the available data instead.

It is important to reiterate that the six selected time series features are in no way exhaustive or imperative. Both using a more data-driven approach to the selection of time series features or selecting entirely different aspects to summarize the time series are legitimate options (also see our discussion of multivariate time series features in the discussion section and see Heylen et al., 2016). Our choice seeks to offer a practical toolbox of time-series features that are common and meaningful to psychological research questions and practice but are also easy to extract and interpret a broad range of developments without asserting strict assumptions.

It also bears repeating, that while our approach allows users to include VAR parameters as one of the possible time series features, it is important to recognize that many contemporary clustering methods focus exclusively on VAR parameters. For instance, packages like clusterVAR (Ernst et al., 2021) and gimme (Gates et al., 2017) rely on these parameters to discern groupings within time series data. Similarly, graphicalVAR (Park et al., 2024) applies VAR-based techniques for clustering in psychological networks. Our feature-based approach subsumes these methods by allowing users to integrate VAR parameters alongside other dynamic features, offering a more comprehensive and flexible framework for clustering.

Feature reduction

Once a meaningful set of time series features has been extracted for each variable and participant, the total number of data points sometimes remains too large for the desired clustering algorithm. As an example, a relatively common scenario would include 10 variables of interest, where eight time series features are extracted, resulting in 80 features per participant (with a common sample size of 100 participants, which would result in a total of 8,000 data points in this hypothetical example). We offer an illustration of the compounding of the numbers of data points in Figure 1. The difficulty of finding stable clusters for data with a large number of dimensions is sometimes termed the 'dimensionality curse' (e.g., Altman & Krzywinski, 2018).

To deal with this dimensionality issue, two main approaches have been proposed—feature selection and feature projection (e.g., Erdogmus et al., 2008). While feature selection refers to the process of identifying and selecting a subset of relevant features from the original feature set (Alelyani et al., 2014), feature projection refers to the process of transforming the original feature set into a new feature set of lower dimensionality (Carreira-Perpiñán, 1997). In general, feature selection procedures have the benefit that they retain the interpretable feature labels directly and immediately indicate which features were most informative in the sample. Feature projection methods, on the other hand, have been popular because they are efficient, widely available, and applicable to a wide range of data types. We provide an overview of common approaches, an intuitive introduction to common methods, and exemplar algorithms in Supplemental Material C.

It is important to note that the necessity and utility of feature reduction depend heavily on the specific clustering algorithm used. Algorithms like k-means, which rely on calculating distances between data points, often struggle with high-dimensional data due to the "curse of dimensionality." In high dimensions, distance measures become less effective, making it difficult for k-means to identify meaningful clusters (Altman & Krzywinski, 2018). Conversely, algorithms like Walktrap, which operate on similarity measures derived from correlation matrices, can actually benefit from higher-dimensional data because more features lead to more robust and accurate similarity estimates between participants (e.g., Gates et al., 2016; Golino & Epskamp, 2017). This abundance of features enhances the algorithm's ability to detect meaningful clusters, improving the reliability of the clustering results. Thus, the decision to reduce features should align with the chosen clustering algorithm and its capacity to handle or leverage high-dimensional data.

For our own illustration data, we chose a feature projection method to reduce the dimensionality of our extracted features. We particularly chose the feature projection method for its broad applicability. We, specifically, selected the commonly used *principal component analysis* (PCA). Some of the more tailor-made feature selection algorithms can be more accurate in reducing the feature dimensionality and might retain feature importance information more directly, depending on the specific data structure. However, PCAs have the distinct benefit that they are well-established within the psychometric literature (Jolliffe, 2011) and can be broadly applied to a wide variety of studies in an automatized manner (Abdi & Williams, 2010). As our aim is to present a general illustration that can also be adopted across use cases, we present the workflow using a PCA here, but we encourage users to consider more specialized methods as well (we provide an example decision guide in Supplemental Material C).

To use the PCA with our extracted time series features, we first standardize all features across participants to ensure that all features are weighted equally (Horne et al., 2020). We then enter all 72 features into the analysis. The PCA uses linear transformations in such a way that the first component captures the most possible variance of the original data (e.g., by finding a vector that maximizes the sum of squared distances Abdi & Williams, 2010; Jolliffe, 2002). The following components will then use the same method to iteratively explain the most of the remaining variance while also ensuring that the components are linearly uncorrelated (Shlens, 2014). In practice, this meant that the PCA decomposed the 72 features into 72 principal components but now (because of the uncorrelated linear transformations) the first few

principal components will capture a majority of the variance. We can then decide how much information (i.e., variance) we are willing to sacrifice for a reduced dimensionality. A common rule of thumb is to use the principal components that jointly explain 70–90% of the original variance (i.e., cumulative percentage explained variance; e.g., Jackson, 2003). For our illustration, we select the first 27 principal components that explain 80% of the variance in the original 72 features (reducing the dimensionality by 62.50%). For the extracted principal components we save the 27 principal component scores for each participant (i.e., the participants' coordinates in the reduced dimensional space; PC-scores).

We would like to comment on two practical matters when using principal components—the amount of dimensionality reduction and the interpretation of the principal components. Regarding the expected dimensionality reduction, given its methodology, PCAs tend to 'work better' at reducing dimensions with (highly) correlated variables (e.g., Jolliffe, 2002). Thus, with a set of very homogeneous variables and features, users will need fewer principal components to explain a large amount of variance, while a more diverse set of variables and features will tend to require more principal components to capture the same amount of variance (e.g., Abdi & Williams, 2010). Our 27 principal components are still a relatively high number of variables, but this is not surprising as we chose a diverse conceptualization and a diverse set of time series features. In terms of interpretability, PCA allows users to extract information on the meaning of the principal components. In particular, because the principal components are linear combinations of the original features, users can extract the relative importance of each feature for the extracted principal components (i.e., the eigenvectors). While this can be useful in understanding the variance in the original data or help with manual feature selection, we use the PCA here purely to reduce the dimensionality for the clustering step. Instead of relying on the principal components, we used the original features of interest to interpret the later extracted clusters. We particularly advocate for such an approach if all original features are considered meaningful in understanding the time series and users would like to retain the features for interpretation (irrespective of the features' importance).

Feature clustering

For the actual clustering of the time-series features, the main aim is to organize participants into groups so that the features of participants within a group are as similar as possible, while the features of people in different groups are as different as possible (Liao, 2005). The crux of clustering is, thus, to have clearly defined and effective measurements of (dis)similarity. Most of the clustering algorithms used today use some form of distance measurement to optimize group assignment (or similarity measurement for qualitative features; see Aghabozorgi et al., 2015). While others have produced excellent overviews of the many clustering approaches available (e.g., Xu & Tian, 2015), the more readily available approaches suitable for most time series feature data can, broadly speaking, be categorized as based on (1) centroids, (2) distributions, (3) density, (4) hierarchies, or (5) a combination thereof (see Supplemental Material C for an overview; also see Jain et al., 1999, for a broader review).

There is, unfortunately, no one-size-fits-all solution to clustering, and users will usually have to make an informed decision based on the structure of their data as well as an appropriate weighing of accuracy and efficiency. We provide a short intuitive explanation for common approaches, together with some of their characteristics and example algorithms in Supplemental Material C. For our own illustration, we have chosen centroid-based k-means clustering. Although k-means sacrifices some level of accuracy, it offers certain advantages. We specifically chose k-means because it is an extremely efficient method that works well with large participant- and feature numbers without making too many restrictive assumptions about the shape of the clusters (Jain, 2010). K-means is also well established within the research community and has been readily implemented in many statistical software packages (Hand & Krzanowski, 2005). Additionally, many of the feature selection methods have been specifically designed for the well-established k-means algorithm (e.g., Boutsidis et al., 2010). As such, the k-means offers a good starting point for many psychological researchers, and the method should be generalizable across a relatively wide variety of projects.

During the k-means clustering itself, the analysis seeks to minimize the total within-cluster variation. The analysis is designed to optimize the clustering of the feature data into k groups, where k is a predefined number of clusters. We used the Hartigan and Wong algorithm, which is a widely used algorithm in k-means clustering (Hartigan & Wong, 1979). The algorithm starts by randomly separating the data points into k clusters and then iteratively updates the

assignment of each point to the nearest cluster center until convergence. To do so, the Hartigan and Wong algorithm specifically calculates the within-cluster variation (W) of cluster C_i as the summed squared Euclidean distances of the feature x to the nearest cluster centroid μ_i :

$$W(C_i) = \sum_{x \in C_i} (\boldsymbol{x} - \boldsymbol{\mu_i})^2 \tag{8}$$

By summing the within-cluster sum of squares from all k clusters, we can then derive the total within-cluster sum of square *WCSS*:

$$WCSS = \sum_{i=1}^{k} W(C_i) = \sum_{i=1}^{k} \sum_{x \in C_i} (\boldsymbol{x} - \boldsymbol{\mu_i})^2 \tag{9}$$

It is this *WCSS* that becomes the objective function to be minimized by iteratively moving features from one cluster to another (Hartigan & Wong, 1979). In particular, the algorithm (1) calculates the cluster centroids of the initial partitioning, (2) checks whether any feature has a centroid that is closer than that of the currently assigned cluster (3) updates the centroids based on any reassigned features, and then iterates between steps two and three until *WCSS* is minimized (i.e., locally optimal convergence) or a maximum number of iterations is reached (Jain, 2010). Given the iterative nature of the algorithm, initial partitioning is often important because the algorithm can arrive at a suboptimal clustering where the *WCSS* cannot be further reduced by moving any feature to another cluster, despite a better solution existing (i.e., a local minimum; Timmerman et al., 2013). It is, therefore, often recommended to run the k-means clustering with several different starting positions.

In our case, we entered the participants' PC-scores from the feature reduction step into the k-means algorithm. Because we did not know the underlying number of clusters within our sample, we calculated the cluster solutions for $k = \{2, \ldots, 10\}$. To avoid local minima, we used 100 random initial centroid positions for each run. Each of the 9 cluster solutions converged within the iteration limit. In the next step, we will then evaluate which of the extracted cluster solutions offers the best fit with the data.

Cluster evaluation

Now that the participants have been assigned to their respective clusters based on the similarity of their time series features, the final evaluation step includes two main elements, (1) evaluating the performance of the clustering analyses to choose an optimal solution and (2) interpreting the extracted clusters conceptually.

Performance

Performance evaluation often means assessing the accuracy, stability, and separation or purity of the clustering (Keogh & Kasetty, 2003). Importantly, any evaluation of the results depends on the research questions, the data, and the methods used. However, broadly speaking, evaluation methods can be categorized based on whether the true cluster labels are known or not (Saxena et al., 2017). If true class labels are known, cluster assignments can be compared to true class labels—using measures such as the F-measure, adjusted Rand index, mutual information and normalized mutual information (i.e., external evaluation; e.g., Liao, 2005). However, if the true cluster assignments are unknown, as with our psychological time series, the quality of the clusters is assessed based on the characteristics of the data itself, such as separation and homogeneity of the clusters, or goodness of fit indices (i.e., internal evaluation; e.g., Aghabozorgi et al., 2015).

In our own illustration example, we used the cluster.stats() function from the fpc R package, which calculates a wide variety of internal cluster validity statistics for each of the extracted clustering solutions. With real-world data, it is not likely that any one evaluation measure will be perfect. Different measures can produce varying results depending on the characteristics of the data and the research question at hand (Kittler et al., 1998). It is, therefore, important to consider a variety of evaluation measures and to carefully interpret the results in the context of a specific analysis (Vinh et al., 2009). We found that across most indices, the analysis with $k = 2$ clusters performed the best. Three commonly reported indices we would like to highlight are the comparison of within-clusters sum of squares, the average silhouette score, and the Calinski-Harabasz index. The first statistic we looked at was the total within-cluster sum of square *WCSS* [see also Equation (9)]. While the within-cluster variation will naturally decrease with (more) smaller clusters, we observed that the decrease in *WCSS* was highest until $k = 2$, after which the decrease was much smaller. This method is also sometimes referred to as the 'elbow method' (Syakur et al., 2018). We then looked at a second, commonly used measure, the average silhouette score. This statistic measures the degree to which each time feature data point is similar to other points within the same cluster, compared to points in other clusters (Rousseeuw, 1987). In our

case, the $k = 2$ solution maximized the silhouette coefficient ($s_2 = 0.09$). Finally, the Calinski-Harabasz index assesses the compactness and separation of the clusters by assessing the ratio of the sum of between-cluster dispersion and of intra-cluster dispersion for all clusters. Thus, a higher score indicates better performance (Calinski & Harabasz, 1974). In our case, the $k = 2$ solution also showed the highest Calinski-Harabasz index ($CH_2 = 16.38$; a full table of all extracted validity statistics is available in Supplemental Material A).[5] In the final $k = 2$ solution the k-means analysis also assigned a relatively even number of participants to cluster 1 ($n_{C_1} = 76$) and cluster 2 ($n_{C_1} = 80$).

To ensure that clustering is necessary in the first place, we also compare the performance to a single-cluster solution (i.e., a single centroid). The comparison with this $k = 1$ solution is slightly different because metrics such as the between-cluster separation are not available. Nonetheless, comparing the within-cluster sums of squares (SS) and the explained variance, we find that two clusters indeed outperform a single cluster solution. Specifically, the total within-cluster SS decreased from 8940.21 for one cluster to 8080.67 for two clusters. Additionally, the variance explained increased from < 0.001 to 0.096 when the cluster count increased to two (e.g., Beijers et al., 2022, ; also see Supplemental Material A for full results).

Interpretation

The interpretation of feature-based time series clustering in psychology involves understanding the meaning and implications of the obtained clusters. In order to make sense of the clustering results, we here focus on three general aspects of the results (Kaufman & Rousseeuw, 1990). (1) Assessing differences between the clusters in the original time series features, (2) comparing the clusters based on prototype developments, (3) comparing the clusters based on between-person differences that were not included in the initial clustering.

In short, we find that the feature-based clustering discerned two meaningfully different groups of participants. We find an adaptive group (cluster 1) that reports higher well-being (*median*: *difference* = −0.52, $t(153.87) = -3.34$, $p = 0.001$, *95%CI* [-0.82, −0.21]; also see Figure 3A) and more positive outgroup interactions (*median*: *difference* = −1.38, $t(152.31) = -11.94$, $p < 0.001$, *95%CI* [-1.61, −1.15]), which are also stable over time (*MAC*: *difference* = 0.54, $t(153.98) = 3.49$, $p < 0.001$, *95%CI* [0.23, 0.84]) and tend to increase more over the 30 day test period (*linear trend*: *difference* = −0.55, $t(149.90) = -3.55$, $p < 0.001$, *95%CI* [-0.85, −0.24]; also see Figure 3C). This group also reported consistently more meaningful (*median*: *difference* = −1.00, $t(136.40) = -7.16$, $p < 0.001$, *95%CI* [-1.28, −0.73]), need-fulfilling (*median*: *difference* = −0.99, $t(135.30) = -7.17$, $p < 0.001$, *95%CI* [-1.26, −0.72]), and cooperative outgroup interactions (*median*: *difference* = −1.33, $t(120.36) = -11.28$, $p < 0.001$, *95%CI* [-1.56, −1.10]). This group with overwhelmingly positive experiences stands in contrast to a more detrimental group (cluster 2). On average, this group reported much less positive, less meaningful, and less fulfilling interactions and interaction patterns (*median*). This group also reported less positive outgroup attitudes, lower well-being and more discrimination experiences (*median*). At the same time, for members of this detrimental cluster (cluster 2) conditions seemed to deteriorate over time (*linear trend*), and there was generally less consistency in the experiences they were able to have (*MAC*, *MAD*, *edf*; also see Figure 3; for a full and interactive comparison of all features see Supplemental Material A).

To identify these patterns, we first inspect the clusters based on the average values of meaningful features (see Figure 2A; Kennedy et al., 2021). We see that for some variables the features are generally stronger in separating the clusters. We, for example, see that the item on '*how cooperative the interaction was*' distinguishes the two clusters across almost all seven features (except for the *auto-correlation*, see Figure 2A). Compare this to the '*outgroup attitudes*' item where the differences between the clusters are much smaller for almost all features. We then inspect the clusters with a focus on the features (see Figure 2B). Although these are the same data as for the variable focus, we can see more clearly that some features are better at distinguishing the clusters across variables. For example, *MAD* and *median* distinguish the two clusters on almost all variables (except for the item of whether the interaction was representative of the outgroup). These two features stand in stark contrast to other features, such as the *lag-1 auto correlations* or the *GAM edf*, which showed much smaller differences between the two clusters (see Figures 2B

[5]It is important to note that another commonly assessed aspect of the evaluation is determining the stability and robustness of the clusters (Berkhin, 2006). This can be assessed by evaluating the sensitivity of the clusters to different feature sets or clustering algorithms, or by using techniques such as bootstrapping to assess the uncertainty of the clusters (Vinh et al., 2009). Especially when comparing different clustering algorithms, a common index is the Bayesian information criterion (BIC), where a lower BIC indicates that a model is more representative of the data (van de Schoot et al., 2017).

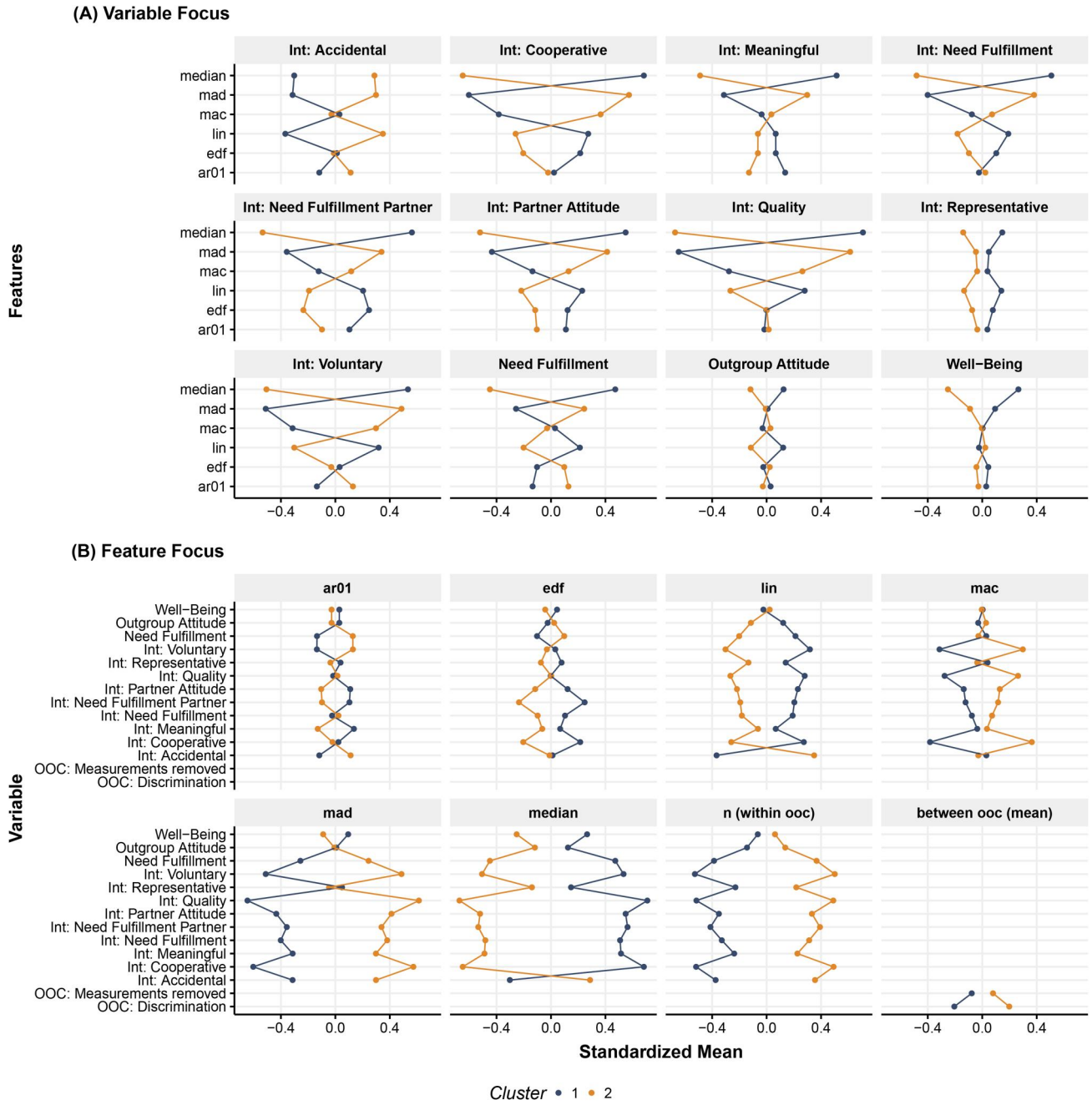

Figure 2. Cluster group comparisons based on features and variables.
Note: "Int." = outgroup interaction, "mad" = median absolute deviation, "mac" = mean absolute change, "lin" = linear slope, "edf" = estimated degrees of freedom of an empty GAM spline model, "ar01" = lag-1 autocorrelation, "OCC"/"occ" = out-of-cluster comparison
Within the "(B) Feature Focus" subplot, the 'n (within ooc)' is an out-of-cluster comparison of the within-person available measurements for each variable; the 'between ooc (mean)' are also out-of-cluster comparisons but on a between-person level. 'Measurements removed' is the person-specific count of measurement occasions removed during the missingness handling and 'Discrimination' is the scale mean of daily discimination experiences (measured during the final survey).

and 3; please note that we offer readers an interactive tool to assess the cluster differences for all features in Supplemental Material A). This offers some information on which features were most important in differentiating the two extracted groups, but also shows that with real-world data, not all features will have enough range to distinguish people on all variables (e.g., see the nonlinearity patterns in Figure 3; for a more direct illustration of GAM edf differences, see Bringmann et al., 2017).

Taking these two perspectives together, we can also focus on individual features or variables, in particular.

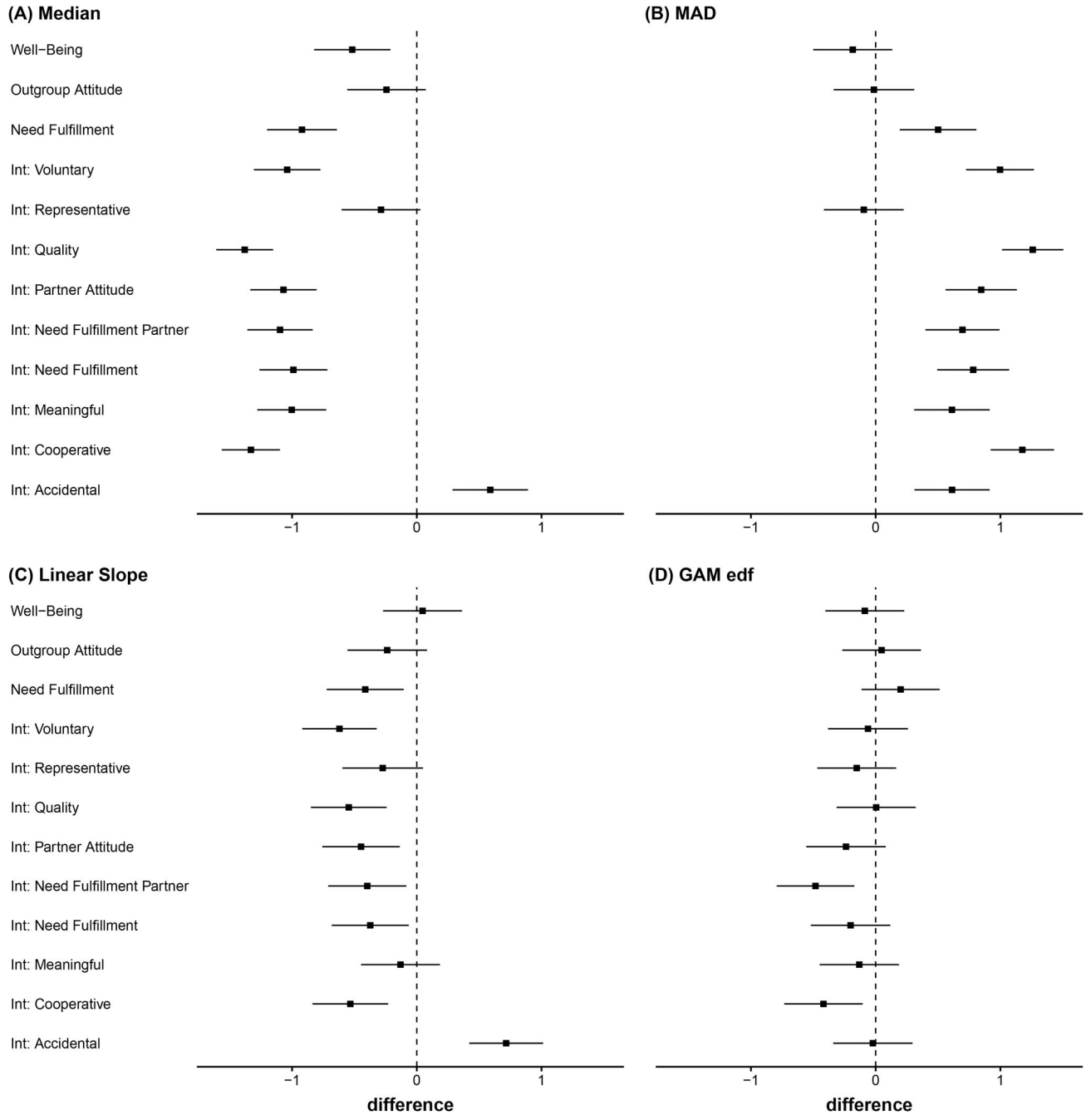

Figure 3. Comparison cluster differences by features and variables.
Note: The figure shows the differences between the clusters in the standardized features that were entered into the dimensionality reduction (for each input variable). We display the median (panel A), the median absolute deviation (MAD, panel B), the univariate linear slope (panel C), as well as the estimated degrees of freedom of the generalized additive model splines (GAM edf, panel D). Please also note that as part of Supplemental Material A, we provide readers with an interactive selection tool to compare cluster differences on all variables and features.

We, for example, see a strong difference in average well-being, where participants in cluster 2 showed a much lower median well-being over the time series (*difference* = −0.52, $t(153.87) = -3.34$, $p = 0.001$, *95%CI* [-0.82, −0.21]). At the same time, in terms of well-being stability, both groups have virtually identical average *MAC* statistics for well-being (*difference* = −0.01, $t(153.96) = -0.04$, $p = 0.968$, *95%CI* [-0.32, 0.31]; also see Figure 2A). There are, thus, variables and features that distinguish the clusters better than others, and a combination of variables and features lets us explore meaningful group differences in more detail. In our case, we see that the central tendency, variability, and linear trend are best at

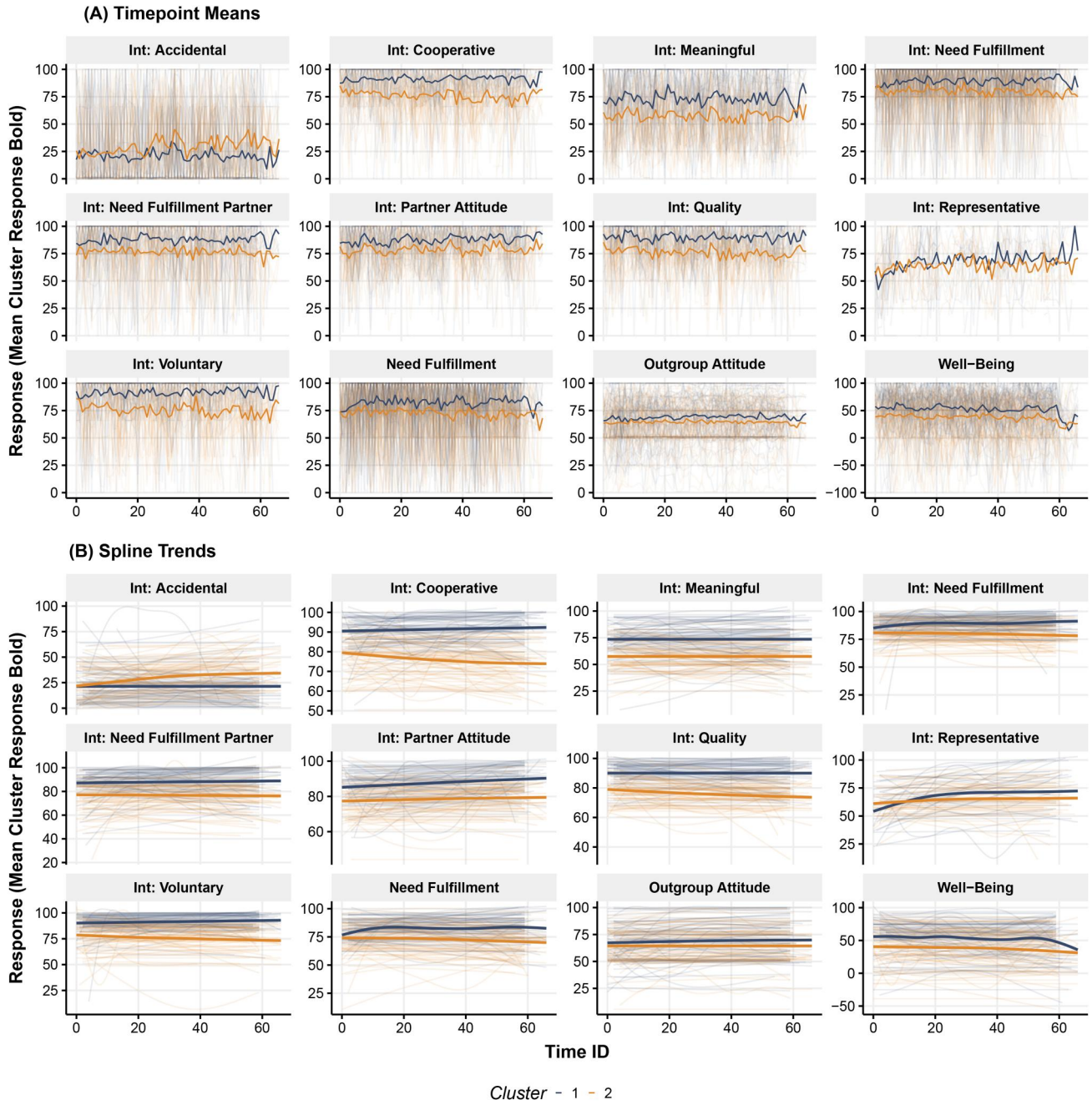

Figure 4. Cluster group comparisons over time.
Note: Subplot (A) displays the variable cluster means at every measurement occasion. The thinner lines represent all individual time series. Subplot (B) shows the GAM spline for each cluster across the measurement occasions. The thinner lines present all individual GAM Splines.

distinguishing a group with mainly positive experiences (cluster 1) from a group with a more negative experience (cluster 2). We also see that our clusters line up with the past literature on the importance of focusing on simpler and more meaningful statistics (Bringmann & Eronen, 2018; Dejonckheere et al., 2019; Eronen & Bringmann, 2021).

In the second step, we look at the prototypical trajectories of the clusters. For k-means clustering it is often recommended to use the average over time of the responses within the cluster (see Figure 4; Niennattrakul & Ratanamahatana, 2007).[6] Immediately striking are the mean differences, where participants in cluster 1 had more meaningful and fulfilling outgroup interactions and also consistently reported more voluntary and cooperative interactions,

[6]It is important to note, however, that direct comparability can be a concern, and often times some subset selection or nonlinear alignment is necessary (e.g., Gupta et al., 1996).

but fewer accidental and involuntary interactions. The same cluster (cluster 1) also reported an increase in need-fulfilling interactions over the 30-day period and an increase in interactions that were representative of the outgroup. Whereas the other cluster (cluster 2) showed a decrease in voluntary, cooperative, and positive interactions over the 30 days. This 'deterioration' cluster (cluster 2) also saw a decrease in general need fulfillment but did not experience well-being over 30 days (see Figure 3C). We also see that while interaction representativeness, outgroup attitudes and well-being are relatively stable for both clusters, the deteriorating cluster (cluster 2) also showed substantially higher variability and instability on most of the other variables (although these effects are much smaller; see Figure 4A).

Finally, we can also assess the clusters across other individual difference variables (e.g., Monden et al., 2022). This out-of-feature comparison allows us to check for data artifacts, as well as to check whether the developmental clusters are associated with important social markers and individual differences. To illustrate artifact checks, we added the number of ESM measurements into the comparison and find that participants in the deterioration cluster (cluster 2) on average completed slightly more ESM surveys in general and reported on more intergroup interactions in particular (see *n* in Figure 2B). In our data exclusion procedures, we ensured that the general time frame and completion rates are similar for all participants, and indeed the numbers in ESM measurements generally are largely similar (e.g., see *n* for well-being and outgroup attitudes). However, the difference in the reported number of interactions might indicate either a clustering artifact or a meaningful difference. The higher average number of interactions in cluster 2 could, for example, indicate a clustering artifact if the variances are substantially larger due to the larger samples (e.g., restriction of range in the smaller sample Kogan et al., 2006). In our case, this seems less likely because one out of four variables did not differ in terms of the MAD (i.e., our selected measurement of the time series variance; see Figure 3 for an illustration). At the same time, however, the difference in the number of experienced interactions might also indicate a meaningful difference, where the deteriorating cluster (cluster 2) on average reported more outgroup interactions ($difference = 1.03$, $t(150.83) = 7.50$, $p < 0.001$, *95%CI* [0.76, 1.30]), but these interactions were less voluntary ($difference = -1.04$, $t(108.89) = -7.71$, $p < 0.001$, *95%CI* [-1.31, −0.77]), less meaningful ($difference = -1.00$, $t(136.40) = -7.16$, $p < 0.001$, *95%CI* [-1.28, −0.73]), and less positive ($difference = -1.38$, $t(152.31) = -11.94$, $p < 0.001$, *95%CI* [-1.61, −1.15]). Thus, while more research is needed for a conclusive test, our data seem to suggest that the differences in reported interactions are a meaningful difference between the clusters. Such a finding would also be in line with past research highlighting the role of negative intergroup interactions in explaining intergroup relations (e.g., Barlow et al., 2012; Graf et al., 2014; Prati et al., 2021). A related validity check was the inclusion of missingness handling, where we compared the two clusters on the average number of measurements removed as part of the missingness handling. We find that the clusters did not differ significantly in this metric, suggesting that missingness handling did not affect the cluster separation (also see Appendix B and Supplemental Material A).

To further illustrate the utility of assessing out-of-feature individual differences, we also compared the two samples in terms of the participants' self-reported discrimination experiences in the Netherlands (measured during the post-measurement). When looking at the group comparison, we find that participants in the deteriorating cluster (cluster 2) reported substantially higher levels of everyday discrimination ($difference = 0.40$, $t(151.71) = 2.56$, $p = 0.011$, *95%CI* [0.09, 0.71]; Figure 2B). Thus, both intensive longitudinal (e.g., the sum of specific ESM measurements) and cross-sectional variables (e.g., general discrimination differences) that were not included in the original clustering step can be used to explore and understand the cluster differences in more detail.

The cluster separation then has a number of empirical and practical applications. First, the clusters are descriptive. With tens of variables, hundreds of participants and thousands of measurements, singular descriptive statistics are often not able to capture the complex patterns that describe the data set. The feature-based clustering offers some direct insight into the complexity within the data set. In our empirical example, we, for example, see that participants are meaningfully distinguished by a combination of high (vs. low) central tendency, variability, and linear trend. Second, the clusters identify important groups. The adaptive and deteriorating groups offer starting points for empirical exploration as well as practical interventions. Researchers can start to explore what exactly distinguishes the two groups further and generate new bottom-up hypotheses. Practitioners in the field of resettlement can use group separation to identify people in need of assistance and can explore contextual factors that could contribute to the difficulties

that some might face. In our illustration, we, for example, found that participants in the deteriorating cluster (cluster 2) reported less need fulfilling interactions over time. Third, the feature-based approach is flexible and meaningful. We were able to use a wide range of time series features that have been central in the ESM literature and were able to use them directly to identify meaningful groups. For our empirical illustration, we, among others, for example, chose to focus on whether participants differed in their average well-being (i.e., *median*), how much their well-being would vary over time (i.e, *MAD*), and whether their well-being would, on average, increase or decrease over time (i.e., *linear trend*). Alternatively, for others cyclical patterns might be more important—for example, whether well-being was higher on weekends. Importantly, in any case, we did not need to translate these dynamic features into probabilistic inference models (e.g., VAR models) to cluster the participants.

To confirm the reliability and thoroughness of our cluster analysis, we undertook several supplementary analyses. These include evaluating the effects of how we managed missing data, examining a reduced model that excludes dynamic characteristics, and providing an enhanced user interface for the investigation of different algorithms. Comprehensive information on these analyses is documented in Appendix B (and full results are available as part of Supplemental Material A). In short, we find that the methods we used are largely robust to different missingness handling decisions, share a reasonable similarity to a more simplified model, and perform consistently well with different parameter options (see Appendix B).

Discussion

The purpose of this article was to introduce feature-based time series clustering as an amenable and transparent approach to understanding between-person differences in developmental patterns of psychological time series data. Rather than relying on person-specific model parameters, which can be restrictive and assumption-bound, we argue for the more flexible and theoretically grounded approach of directly clustering on relevant features of the time-series data. By leveraging the rich array of dynamic measures, the approach offers the advantages of flexibility, few strict assumptions, and high interpretability, thus potentially enriching our understanding of heterogeneous psychological processes in intensive longitudinal studies.

To illustrate the practical utility of the approach, we applied the method to empirical data from the real world that highlight common ESM issues of multivariate conceptualizations, structural missingness, and nonlinear trends (e.g., Ariens et al., 2020). With the real-world data, we followed a stepwise approach to discuss key issues during input selection, feature extraction, feature reduction, feature clustering, and cluster evaluation. Within this stepwise approach, our article shows that feature-based clustering offers a meaningful fit for psychological research, as both the time-series features and the analysis steps are well established within the field, and statistical packages are readily available. Time series features (such as means or linear trends) are not only easy to extract, but also hold conceptual meaning for psychological data and can be chosen to address specific research questions (also see Table 3).

Importantly, we show that feature-based clustering is not only approachable but provides interpretable and transparent insights about the grouped patterns. For our example of migration experiences, the method was useful to discern adaptive from more stressful experiences and helped to contextualize divergent experiences. We found that some variables, such as perceptions of the quality of the interaction or the fulfillment of the needs, were particularly important in distinguishing the groups (see Figure 2A). Similarly, we found that the central tendency (*median*), variability (*MAD*), and linear trend (*slope*) were the most impactful dynamic features in discerning the trajectory clusters (this is further emphasized by a simpler model using only *median* and *MAD* performing similarly well, see Appendix B. Also see Figure 2B). Jointly, these two approaches allowed us to identify a cluster that had generally positive and improving experiences, while the other cluster had more negative and deteriorating experiences. We were even able to further contextualize the results with out-of-feature comparisons, where we found that the group with the more difficult experiences also reported substantially more discrimination experiences during the post-test (see, e.g., Figure 2B). In summary, the feature-based approach enables us to identify directly interpretable and meaningful groups, providing transparency regarding the data input on which the clusters are based.

Before we turn to the formal limitations of the feature-based clustering approach, we would like to briefly address the role of multivariate time series features. Multivariate features are those that capture contemporaneous or dynamic relationships between the different time series within a person Kuppens and Verduyn (2017). These features can include average correlations and co-variances or the cross-lagged

correlation equivalents, as well as parameters that are based on these (lagged) multivariate relationships (e.g., VAR parameters). Recently, cross-lagged effects have also been extracted from dynamic network models (e.g., see Wendt et al., 2020). We have chosen not to include multivariate features within this illustration. We have done so mainly because these features often add a much larger computational load to the model. As an example, for our example set of 12 variables, even a simple lag-1 VAR model would add 156 additional features (12 variables $\times$ (12 lagged parameters $+$ 1 intercept)). That are more than twice as many features per person than the six univariate features we selected combined ($6 \times 12 = 72$). While the feature-based approach can technically handle the parameters and the dimensionality reduction can deal with the added number of dimensions, we seek to introduce the method with approachable dimensionality reduction and clustering models, which are not ideal for such a large number of input features (e.g., assessing over 150 additional parameter differences during the cluster interpretation). Alternative dimensionality reduction approaches would make the interpretation more straightforward, but their use is often much more specialized and bound to specific cluster models (e.g., methods that select the most influential features instead of projecting to a lower-dimensional space, see Supplemental Material C).[7]

It is important to note here though that the method can directly accommodate multivariate features and that such features are commonly of interest within the literature. Particularly when theory testing models are developed as part of the research process, already adding the parameters as time series features often fits naturally within the research cycle. However, given the cautionary remarks here, we recommend a careful use of model parameters in combination with other time series features (i.e., either a more selective model-building process where not all variables are included or a filter process). It should also be noted that the feature-based clustering approach is inherently a multivariate process in that the model takes into account the features of variables and considers them jointly.

Limitations

While feature-based time-series clustering offers a promising approach to understanding psychological time-series data, it is not without limitations. In particular, feature-based clustering has both usability- and robustness limitations across its multiple steps.

In terms of convenience, each of the steps requires users to make an informed decision about the choice of method and algorithm. These additional steps of decision-making and transparency increase the initial barrier to entry. We hope that our empirical illustration, the sample code, and the custom functions, offer a relatively generalizable procedure that showcases the ease of use, but clustering, unfortunately, does not offer a universal one-size-fits-all solution.

In terms of methodological robustness, the variety of methods in each of the steps also brings with it the potential inconsistent results between methods (e.g., Bastiaansen et al., 2019). A different set of variables, time series features, or a different clustering algorithm might have resulted in substantially different cluster assignments. While the variety and diversity of methods are helpful in finding options even for more complex types of data, different algorithms often offer different results (e.g., Keogh & Lin, 2005). And even when patterns produce robust clustering solutions across algorithms, individual methods might still have their idiosyncratic shortcomings (Xu & Tian, 2015).

As an illustration, the choice of time-series features to extract from the time-series data is a critical step that can significantly influence the results of the clustering process. In the current example, we chose to extract time series features such as medians, autocorrelations, and linear trends, which are psychologically and conceptually meaningful in interpreting our time series clusters. However, this selection is not exhaustive and may not capture all relevant aspects of the time series data. For example, we did not consider attributes like periodicity or spectral density, which could shed light on the cyclical patterns of the data. The choice of time-series features largely hinges on the researcher's specific research question and assumptions about the data, thereby injecting a level of subjectivity into the process. Similar challenges arise with the choice of the clustering algorithm or the cluster illustration. These challenges are not unique to feature-based clustering, rather they are common to all clustering approaches (Horne et al., 2020; Liao, 2005). However, it is important to remember that multi-step data-driven approaches are particularly vulnerable to the impact of the researchers' degrees of freedom (Beijers et al., 2022). Additionally, during our additional analyses (see Appendix B), when we evaluated the simplified model (only including central tendency and variability), the cluster

[7]Please note that this issue would be less prevalent with more global measurements such as interdependence measures (e.g., network density) or diversity measurements (e.g., Gini coefficient). Thesemeasures are however also not without criticism (Bringmann et al., 2019).

results were very similar to those of the main analysis. While there were some nuances to this finding, the results highlight that the use of complex dynamic parameters should be directly linked to the research question and should only be added if the dynamic patterns are to be expected within a given data set.

One potential remedy to many of the limitations of feature-based clustering lies in transparent and reproducible reporting of the user's decisions for each of the analysis steps. In our own description of the method, we have provided a range of options and have motivated our own choices to facilitate the transparency of the individual steps and decision moments. We have additionally created Supplemental Material C to illustrate the decision process of different methods and offer Supplemental Material B to explore different options. Beyond the structures and resources provided here, van de Schoot et al. (2017) have proposed an extensive checklist for latent trajectory studies. Most of their recommendations and reporting guidelines also apply to feature-based clustering and might even offer a template for researchers who want to pre-register their analysis procedures (also see Kirtley et al., 2021).

Adding to the discussion on methodological limitations, it is also crucial to consider the impact of data granularity on the classification accuracy of feature-based time series clustering. The effectiveness of clustering algorithms is not only contingent on the choice of variables and algorithms, but also significantly influenced by the number and quality of data points per individual. Research indicates that longer time series can enhance the clustering outcome by providing a more detailed view of the underlying patterns (Liao, 2005; Montero & Vilar, 2014). However, the minimum number of data points required for accurate classification remains a subject of ongoing investigation, and existing studies suggest that this threshold may vary depending on the complexity of the data and the features used in the analysis (Aghabozorgi et al., 2015). This aspect underscores the necessity for a careful and nuanced approach to feature selection and algorithm application and the continued need for research to optimize the balance between data complexity and clustering accuracy.

Implications

Notwithstanding the limitations, we believe that feature-based clustering offers new potential for researchers and practitioners to assess psychological time series.

For researchers, the feature-based time series clustering approach offers a number of compelling implications. The flexibility and interpretability mean that feature-based time series clustering can be applied to a wide range of data types and research questions. The method can be used to contextualize preexisting groups by extracting their time series features and comparing a data-driven approach with existing group labels. Furthermore, the feature-based approach can also be used as an exploratory, descriptive, or predictive approach to intensive longitudinal data. By reducing the complexities of ESM data to important and meaningful patterns, a bottom-up approach can aid in the creation of more embedded theories and interventions, or simply in describing the often complex and heterogeneous data researchers collect during ESM studies.

Beyond the direct academic use, the feature-based time series clustering approach also addresses practical and applied uses. For practitioners with appropriate training, the approach offers a practical and grounded method for dealing with the challenges of complex and messy data from multiple patients, customers, or users. The approach not only directly deals with dimensionality, missingness, and time scales in the time series, but the interpretability and transparency aspects offer particular utility in applied settings. Additionally, the approach is also more readily accessible to practitioners who may not have extensive training in complex data analysis techniques. We provide practical algorithm overviews and readily available code for data preparation, analysis, and interpretation. The ability to identify and interpret meaningful patterns in time series data can have significant implications for practice, particularly in fields such as clinical, organizational, or social psychology, where understanding individual differences and developmental patterns can inform interventions and decision processes.

In conclusion, we show that feature-based time series clustering can effectively reduce the complexities of psychological time series data to important and meaningful patterns. It does so with more flexibility, versatility, and less strict assumptions than many of the commonly used approaches to date. As such, the feature-based time series clustering approach addresses key challenges in the field and aids researchers and practitioners in describing and exploring patterns across participants. We hope that the method adds to the methodological toolkit of ESM researchers and promotes the creation of more embedded methods, theories, and interventions.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work received no funding.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis: Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Adya, M., Collopy, F., Armstrong, J., & Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting*, *17*(2), 143–157. https://doi.org/10.1016/S0169-2070(01)00079-6

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Alelyani, S., Tang, J., & Liu, H. (2014). Feature selection for clustering: A review. In C. C. Aggarwal & C. K. Reddy (Eds.), *Data clustering: Algorithms and applications* (pp. 29–60). Chapman & Hall/CRC.

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., Liu, H., Schultz, T., & Gamboa, H. (2020). TSFEL: Time series feature extraction library. *SoftwareX*, *11*, 100456. https://doi.org/10.1016/j.softx.2020.100456

Barlow, F. K., Paolini, S., Pedersen, A., Hornsey, M. J., Radke, H. R., Harwood, J., Rubin, M., & Sibley, C. G. (2012). The contact caveat: Negative contact predicts increased prejudice more than positive contact predicts reduced prejudice. *Personality & Social Psychology Bulletin*, *38*(12), 1629–1643. https://doi.org/10.1177/0146167212457953

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., De Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R. G., Oravecz, Z., Riese, H., Rubel, J., & Bringmann, L. F. (2019). Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211. https://doi.org/10.31234/osf.io/c8vp7

Beijers, L., Van Loo, H. M., Romeijn, J.-W., Lamers, F., Schoevers, R. A., & Wardenaar, K. J. (2022). Investigating data-driven biological subtypes of psychiatric disorders using specification-curve analysis. *Psychological Medicine*, *52*(6), 1089–1100. https://doi.org/10.1017/S0033291720002846

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25–71). Springer-Verlag. https://doi.org/10.1007/3-540-28349-8_2

Berndt, D. J., Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *AAAIWS'94: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.5555/3000850.3000887

Berry, J. W. (1986). The acculturation process and refugee behavior. *Refugee Mental Health in Resettlement Countries*, *10*(75), 25–37.

Bertenthal, B. I. (2007). Dynamical systems: It's about time. In S. M. Boker & M. J. Wenger (Eds.), *Data analytic*

techniques for dynamical systems (pp. 1–24). Lawrence Erlbaum Associates.

Boker, S. M., & Graham, J. (1998). A dynamical systems analysis of adolescent substance abuse. *Multivariate Behavioral Research*, *33*(4), 479–507. https://doi.org/10.1207/s15327906mbr3304_3

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Bos, E. H., De Jonge, P., & Cox, R. F. A. (2019). Affective variability in depression: Revisiting the inertia–instability paradox. *British Journal of Psychology (London, England: 1953)*, *110*(4), 814–827. https://doi.org/10.1111/bjop.12372

Boutsidis, C., Drineas, P., & Mahoney, M. W. (2010). Unsupervised feature selection for the k-means clustering problem. In *NIPS'09: Proceedings of the 22nd International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, Curran.

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review*, *125*(4), 606–615. https://doi.org/10.1037/rev0000108

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy*, *149*, 104011. https://doi.org/10.1016/j.brat.2021.104011

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T. W., & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892–903. https://doi.org/10.1037/abn0000446

Bringmann, L. F., Ferrer, E., Hamaker, E. L., Borsboom, D., & Tuerlinckx, F. (2018). Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model. *Multivariate Behavioral Research*, *53*(3), 293–314. https://doi.org/10.1080/00273171.2018.1439722

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*, *22*(3), 409–425. https://doi.org/10.1037/met0000085

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR(1) based models do not always outpredict AR(1) models in typical psychological applications. *Psychological Methods*, *23*(4), 740–756. https://doi.org/10.1037/met0000178

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Clustering vector autoregressive models: Capturing qualitative differences in within-person dynamics. *Frontiers in Psychology*, *7*, 1540. https://doi.org/10.3389/fpsyg.2016.01540

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3) 1049–1064. https://doi.org/10.18637/jss.v045.i03

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Caro-Martín, C. R., Delgado-García, J. M., Gruart, A., & Sánchez-Campusano, R. (2018). Spike sorting based on shape, phase, and distribution features, and K-TOPS clustering with validity and error indices. *Scientific Reports*, *8*(1), 17796. https://doi.org/10.1038/s41598-018-35491-4

Carreira-Perpiñán, M. Á. (1997). *A review of dimension reduction techniques* (Technical Report CS-96-09). Department of Computer Science, University of Sheffield.

Castro-Alvarez, S., Bringmann, L. F., Meijer, R. R., & Tendeiro, J. N. (2024). A time-varying dynamic partial credit model to analyze polytomous and multivariate time series data. *Multivariate Behavioral Research*, *59*(1), 78–97. https://doi.org/10.1080/00273171.2023.2214787

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 67–128). Rand McNally.

Chatfield, C. (2003). *The analysis of time series*. Chapman; Hall/CRC. https://doi.org/10.4324/9780203491683

Choi, J., Miller, A., & Wilbur, J. (2009). Acculturation and depressive symptoms in Korean immigrant women. *Journal of Immigrant and Minority Health*, *11*(1), 13–19. https://doi.org/10.1007/s10903-007-9080-8

Chow, S.-M., Zu, J., Shifren, K., & Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research*, *46*(2), 303–339. https://doi.org/10.1080/00273171.2011.563697

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – A Python package). *Neurocomputing*, *307*, 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS One*, *11*(12), e0167490. https://doi.org/10.1371/journal.pone.0167490

D'Mello, S. K., & Gruber, J. (2021). Emotional regularity: Associations with personality, psychological health, and occupational outcomes. *Cognition & Emotion*, *35*(8), 1460–1478. https://doi.org/10.1080/02699931.2021.1968797

de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology*, *8*, 1849. https://doi.org/10.3389/fpsyg.2017.01849

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

den Teuling, N., Pauws, S., & van den Heuvel, E. (2021). Clustering of longitudinal data: A tutorial on a variety of approaches. *arXiv preprint arXiv:2111.05469*. https://doi.org/10.48550/ARXIV.2111.05469

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic

processes in psychopathology. *Journal of Abnormal Psychology*, *118*(1), 195–202. https://doi.org/10.1037/a0014868

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Erdogmus, D., Ozertem, U., & Lan, T. (2008). Information theoretic feature selection and projection. In J. Kacprzyk, B. Prasad, & S. R. M. Prasanna (Eds.), *Speech, audio, image and biomedical signal processing using neural networks*. Series Title: Studies in Computational Intelligence (pp. 1–22). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75398-8_1

Ernst, A. F., Timmerman, M. E., Jeronimus, B. F., & Albers, C. J. (2021). Insight into individual differences in emotion dynamics with clustering. *Assessment*, *28*(4), 1186–1206. https://doi.org/10.1177/1073191119873714

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*, 100785. https://doi.org/10.1016/j.newideapsych.2020.100785

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. *ACM SIGMOD Record*, *23*(2), 419–429. https://doi.org/10.1145/191843.191925

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781315382722

Fulcher, B. D., Little, M. A., & Jones, N. S. (2013). Highly comparative time-series analysis: The empirical structure of time series and their methods. *Journal of the Royal Society, Interface*, *10*(83), 20130048. https://doi.org/10.1098/rsif.2013.0048

Gates, K. M., Henry, T., Steinley, D., & Fair, D. A. (2016). A Monte Carlo evaluation of weighted community detection algorithms. *Frontiers in Neuroinformatics*, *10*, 45. https://doi.org/10.3389/fninf.2016.00045

Gates, K. M., Lane, S. T., Varangis, E., Giovanello, K., & Guskiewicz, K. (2017). Unsupervised classification during time-series model building. *Multivariate Behavioral Research*, *52*(2), 129–148. https://doi.org/10.1080/00273171.2016.1256187

Glenn, C. R., Kleiman, E. M., Kearns, J. C., Santee, A. C., Esposito, E. C., Conwell, Y., & Alpert-Gillis, L. J. (2022). Feasibility and acceptability of ecological momentary assessment with high-risk suicidal adolescents following acute psychiatric care. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *51*(1), 32–48. https://doi.org/10.1080/15374416.2020.1741377

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Gottman, J. M., McFall, R. M., & Barnett, J. T. (1969). Design and analysis of research using time series. *Psychological Bulletin*, *72*(4), 299–306. https://doi.org/10.1037/h0028021

Graf, S., Paolini, S., & Rubin, M. (2014). Negative intergroup contact is more influential, but positive intergroup contact is more common: Assessing contact prominence and contact prevalence in five Central European countries. *European Journal of Social Psychology*, *44*(6), 536–547. https://doi.org/10.1002/ejsp.2052

Gupta, L., Molfese, D., Tammana, R., & Simos, P. (1996). Nonlinear alignment and averaging for estimating the evoked potential. *IEEE Transactions on Bio-Medical Engineering*, *43*(4), 348–356. https://doi.org/10.1109/10.486255

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hand, D. J., & Krzanowski, W. J. (2005). Optimising k-means clustering results with standard software packages. *Computational Statistics & Data Analysis*, *49*(4), 969–973. https://doi.org/10.1016/j.csda.2004.06.017

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2021). A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research*, *56*(1), 120–149. https://doi.org/10.1080/00273171.2020.1743630

Hayes, A. M., Laurenceau, J.-P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical Psychology Review*, *27*(6), 715–723. https://doi.org/10.1016/j.cpr.2007.01.008

Helmich, M. A., Olthof, M., Oldehinkel, A. J., Wichers, M., Bringmann, L. F., & Smit, A. C. (2021). Early warning signals and critical transitions in psychopathology: Challenges and recommendations. *Current Opinion in Psychology*, *41*, 51–58. https://doi.org/10.1016/j.copsyc.2021.02.008

Helmich, M. A., Wichers, M., Olthof, M., Strunk, G., Aas, B., Aichhorn, W., Schiepek, G., & Snippe, E. (2020). Sudden gains in day-to-day change: Revealing nonlinear patterns of individual improvement in depression. *Journal of Consulting and Clinical Psychology*, *88*(2), 119–127. https://doi.org/10.1037/ccp0000469

Heylen, J., Van Mechelen, I., Verduyn, P., & Ceulemans, E. (2016). KSC-N: Clustering of hierarchical time profile data. *Psychometrika*, *81*(2), 411–433. https://doi.org/10.1007/s11336-014-9433-x

Horne, E., Tibble, H., Sheikh, A., & Tsanas, A. (2020). Challenges of clustering multimodal clinical data: Review of applications in asthma subtyping. *JMIR Medical Informatics*, *8*(5), e16452. https://doi.org/10.2196/16452

Hosenfeld, B., Bos, E. H., Wardenaar, K. J., Conradi, H. J., van der Maas, H. L. J., Visser, I., & de Jonge, P. (2015). Major depressive disorder as a nonlinear dynamic system:

Bimodality in the frequency distribution of depressive symptoms over time. *BMC Psychiatry*, *15*(1), 222. https://doi.org/10.1186/s12888-015-0596-5

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Huang, H.-C., & Jansen, B. (1985). EEG waveform analysis by means of dynamic time-warping. *International Journal of Bio-Medical Computing*, *17*(2), 135–144. https://doi.org/10.1016/0020-7101(85)90084-4

Jackson, J. E. (2003). *A user's guide to principal components*. Wiley-Interscience.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, *13*(4), 354–375. https://doi.org/10.1037/a0014173

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jeronimus, B. F. (2019). Dynamic system perspectives on anxiety and depression. In E. S. Kunnen, N. M. de Ruiter, B. F. Jeronimus, & M. A. van der Gaag (Eds.), *Psychosocial development in adolescence* (1st ed.) (pp. 100–126). Routledge. https://doi.org/10.4324/9781315165844-7

Jolliffe, I. (2002). *Principal component analysis*. Springer.

Jolliffe, I. (2011). Principal component analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1094–1096). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_455

Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). *Finding groups in data*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801

Keil, T. F., Koschate, M., & Levine, M. (2020). Contact logger: Measuring everyday intergroup contact experiences in near-time. *Behavior Research Methods*, *52*(4), 1568–1586. https://doi.org/10.3758/s13428-019-01335-w

Kennedy, B., Reimer, N. K., & Dehghani, M. (2021). Explaining Explainability: Interpretable machine learning for the behavioral sciences. https://doi.org/10.31234/osf.io/9h6qr

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*(4), 349–371. https://doi.org/10.1023/A:1024988512476

Keogh, E., & Lin, J. (2005). Clustering of time-series subsequences is meaningless: Implications for previous and future research. *Knowledge and Information Systems*, *8*(2), 154–177. https://doi.org/10.1007/s10115-004-0172-7

Kim, Y. Y. (2017). Cross-cultural adaptation. In *Oxford research encyclopedia of communication*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.21

Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592092468. https://doi.org/10.1177/2515245920924686

Kittler, J., Hatef, M., Duin, R., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239. https://doi.org/10.1109/34.667881

Kivelä, L., van der Does, W. A. J., Riese, H., Antypa,., & N., 4. (2022). Don't miss the moment: A systematic review of ecological momentary assessment in suicide research. *Frontiers in Digital Health*, *4*, 876595. https://doi.org/10.3389/fdgth.2022.876595

Kiwuwa-Muyingo, S., Oja, H., Walker, S. A., Ilmonen, P., Levin, J., & Todd, J. (2011). Clustering based on adherence data. *Epidemiologic Perspectives & Innovations: EP+I*, *8*(1), 3. https://doi.org/10.1186/1742-5573-8-3

Kogan, J., Nicholas, C. K., & Teboulle, M. (Eds.). (2006). *Grouping multidimensional data: Recent advances in clustering*. Springer.

Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, *13*(6), 1132–1141. https://doi.org/10.1037/a0033579

Kreienkamp, J., Agostini, M., Bringmann, L. F., de Jonge, P., & Epstude, K. (2022). *Psychological needs during intergroup contact: Three experience sampling studies*. Department of Psychology, University of Groningen.

Kreienkamp, J., Agostini, M., Monden, R., Epstude, K., de Jonge, P., & Bringmann, L. F. (2023a). Feature-Based Clustering of Psychological Time Series [OSF repository: Materials, data, code]. https://doi.org/10.17605/OSF.IO/J8DZV

Kreienkamp, J., Agostini, M., Monden, R., Epstude, K., de Jonge, P., & Bringmann, L. F. (2023b). Illustration: Feature Based Time Series Clustering in Psychology. https://janniscodes.github.io/ts-feature-clustering-illustration/

Kreienkamp, J., Agostini, M., Monden, R., Epstude, K., de Jonge, P., & Bringmann, L. F. (2023c). migration-trajectories [GitHub repository: Materials, computer code]. https://janniscodes.github.io/migration-trajectories/

Kreienkamp, J., Agostini, M., Monden, R., Epstude, K., de Jonge, P., & Bringmann, L. F. (2023d). tsFeatureExtracR (Version 0.1.1) [Computer software]. https://doi.org/10.5281/zenodo.8221675

Kreienkamp, J., Agostini, M., Epstude, K., Monden, R., De Jonge, P., & Bringmann, L. F. (2024). tsFeatureClustRApp (Version 1.0.0) [Computer software]. https://github.com/JannisCodes/tsFeatureClustRApp

Kreienkamp, J., Bringmann, L. F., Engler, R. F., de Jonge, P., & Epstude, K. (2024). The migration experience: A conceptual framework and systematic scoping review of psychological acculturation. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *28*(1), 81–116. https://doi.org/10.1177/10888683231183479

Krone, T., Albers, C. J., Kuppens, P., & Timmerman, M. E. (2018). A multivariate statistical model for emotion dynamics. *Emotion*, *18*(5), 739–754. https://doi.org/10.1037/emo0000384

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, *12*(2), 283–289. https://doi.org/10.1037/a0025046

Lacasa, L., Nicosia, V., & Latora, V. (2015). Network structure of multivariate time series. *Scientific Reports*, *5*(1), 15508. https://doi.org/10.1038/srep15508

Lafit, G., Meers, K., & Ceulemans, E. (2022). A systematic study into the factors that affect the predictive accuracy of multilevel VAR(1) models. *Psychometrika*, *87*(2), 432–476. https://doi.org/10.1007/s11336-021-09803-z

Lavori, P. W., Brown, C. H., Duan, N., Gibbons, R. D., & Greenhouse, J. (2008). Missing data in longitudinal clinical trials part A: Design and conceptual issues. *Psychiatric Annals*, *38*(12), 784–792. https://doi.org/10.3928/00485713-20081201-04

Liao, T. W. (2005). Clustering of time series data—A survey. *Pattern Recognition*, *38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.

Loftus, T. J., Shickel, B., Balch, J. A., Tighe, P. J., Abbott, K. L., Fazzone, B., Anderson, E. M., Rozowsky, J., Ozrazgat-Baslanti, T., Ren, Y., Berceli, S. A., Hogan, W. R., Efron, P. A., Moorman, J. R., Rashidi, P., Upchurch, G. R., & Bihorac, A. (2022). Phenotype clustering in health care: A narrative review for clinicians. *Frontiers in Artificial Intelligence*, *5*, 842306. https://doi.org/10.3389/frai.2022.842306

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation [Publisher: Elsevier Inc. *Journal of Clinical Epidemiology*, *110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Maharaj, E. A., D'Urso, P., & Caiado, J. (2019). *Time series clustering and classification* (1st ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9780429058264

Marx, B. D., & Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, *28*(2), 193–209. https://doi.org/10.1016/S0167-9473(98)00033-4

McLean, D. C., Nakamura, J., & Csikszentmihalyi, M. (2017). Explaining system missing: Missing data and experience sampling method. *Social Psychological and Personality Science*, *8*(4), 434–441. https://doi.org/10.1177/1948550617708015

Molenaar, P. C. M., Sinclair, K. O., Rovine, M. J., Ram, N., & Corneal, S. E. (2009). Analyzing developmental processes on an individual level using nonstationary time series modeling. *Developmental Psychology*, *45*(1), 260–271. https://doi.org/10.1037/a0014170

Monden, R., Rosmalen, J. G. M., Wardenaar, K. J., & Creed, F. (2022). Predictors of new onsets of irritable bowel syndrome, chronic fatigue syndrome and fibromyalgia: The lifelines study. *Psychological Medicine*, *52*(1), 112–120. https://doi.org/10.1017/S0033291720001774

Monden, R., Wardenaar, K. J., Stegeman, A., Conradi, H. J., & De Jonge, P. (2015). Simultaneous decomposition of depression heterogeneity on the person-, symptom- and time-level: The use of three-mode principal component analysis. *PloS One*, *10*(7), e0132765. https://doi.org/10.1371/journal.pone.0132765

Montero, P., & Vilar, J. A. (2014). TSclust: An *R* package for time series clustering. *Journal of Statistical Software*, *62*(1), 1–43. https://doi.org/10.18637/jss.v062.i01

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *17*(2), 123–132. https://doi.org/10.1002/wps.20513

Neubauer, A. B., & Schmiedek, F. (2020). Studying within-person variation and within-person couplings in intensive longitudinal data: Lessons learned and to be learned. *Gerontology*, *66*(4), 332–339. https://doi.org/10.1159/000507993

Niennattrakul, V., & Ratanamahatana, C. A. (2007). Inaccuracies of shape averaging method using dynamic time warping for time series data. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, Y. Shi, G. D. van Albada, J. Dongarra, & P. M. A. Sloot (Eds.), *Computational science – ICCS 2007*. Series Title: Lecture Notes in Computer Science (pp. 513–520). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72584-8_68

Nyblom, J. (1986). Testing for deterministic linear trend in time series. *Journal of the American Statistical Association*, *81*(394), 545–549. https://doi.org/10.1080/01621459.1986.10478302

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2016). Bayesian data analysis with the bivariate hierarchical Ornstein-Uhlenbeck process model. *Multivariate Behavioral Research*, *51*(1), 106–119. https://doi.org/10.1080/00273171.2015.1110512

Ou, L., Hunter, M. D., Lu, Z., Stifter, C. A., & Chow, S. (2023). Estimation of nonlinear mixed-effects continuous-time models using the continuous-discrete extended Kalman filter. *The British Journal of Mathematical and Statistical Psychology*, *76*(3), 462–490. https://doi.org/10.1111/bmsp.12318

Park, J. J., Chow, S.-M., Epskamp, S., & Molenaar, P. C. M. (2024). Subgrouping with chain graphical VAR models. *Multivariate Behavioral Research*, *59*(3), 543–565. https://doi.org/10.1080/00273171.2023.2289058

Prati, F., Kana Kenfack, C. S., Koser Akcapar, S., & Rubini, M. (2021). The role of positive and negative contact of migrants with native people in affecting their future interactions. Evidence from Italy and Turkey. *International Journal of Intercultural Relations*, *85*(March), 191–203. https://doi.org/10.1016/j.ijintrel.2021.09.015

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *The Behavioral and Brain Sciences*, *20*(4), 537–556. https://doi.org/10.1017/S0140525X97001581

Ram, N., Conroy, D. E., Pincus, A. L., Lorek, A., Rebar, A., Roche, M. J., Coccia, M., Morack, J., Feldman, J., & Gerstorf, D. (2014). Examining the interplay of processes across multiple time-scales: Illustration with the intraindividual study of affect, health, and interpersonal behavior (iSAHIB). *Research in Human Development*, *11*(2), 142–160. https://doi.org/10.1080/15427609.2014.906739

Räsänen, T., & Kolehmainen, M. (2009). Feature-based clustering for electricity use time series data. In M. Kolehmainen, P. Toivanen, & B. Beliczynski (Eds.), *Adaptive and natural computing algorithms*. Series Title: Lecture Notes in Computer Science (pp. 401–412). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04921-7_41

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Reitsema, A. M., Jeronimus, B. F., van Dijk, M., Ceulemans, E., van Roekel, E., Kuppens, P., & de Jonge, P. (2023). Distinguishing dimensions of emotion dynamics across 12 emotions in adolescents' daily lives. *Emotion*, *23*(6), 1549–1561. https://doi.org/10.1037/emo0001173

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schreuder, M. J., Hartman, C. A., George, S. V., Menne-Lothmann, C., Decoster, J., van Winkel, R., Delespaul, P., De Hert, M., Derom, C., Thiery, E., Rutten, B. P. F., Jacobs, N., van Os, J., Wigman, J. T. W., & Wichers, M. (2020). Early warning signals in psychopathology: What do they tell? *BMC Medicine*, *18*(1), 269. https://doi.org/10.1186/s12916-020-01742-3

Schrodt, P. A., & Gerner, D. J. (2000). Cluster-based early warning indicators for political change in the contemporary levant. *American Political Science Review*, *94*(4), 803–817. https://doi.org/10.2307/2586209

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*. https://doi.org/10.48550/ARXIV.1404.1100

Stefanovic, M., Rosenkranz, T., Ehring, T., Watkins, E. R., & Takano, K. (2022). Is a high association between repetitive negative thinking and negative affect predictive of depressive symptoms? A clustering approach for experience-sampling data. *Clinical Psychological Science*, *10*(1), 74–89. https://doi.org/10.1177/21677026211009495

Suls, J., & Rothman, A. (2004). Evolution of the biopsychosocial model: Prospects and challenges for health psychology. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *23*(2), 119–125. https://doi.org/10.1037/0278-6133.23.2.119

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136. https://doi.org/10.1177/0146167298242002

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, *336*, 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Timm, C., Ubl, B., Zamoscik, V., Ebner-Priemer, U., Reinhard, I., Huffziger, S., Kirsch, P., & Kuehner, C. (2017). Cognitive and affective trait and state factors influencing the long-term symptom course in remitted depressed patients. *PloS One*, *12*(6), e0178759. https://doi.org/10.1371/journal.pone.0178759

Timmerman, M. E., Ceulemans, E., De Roover, K., & Van Leeuwen, K. (2013). Subspace K-means clustering. *Behavior Research Methods*, *45*(4), 1011–1023. https://doi.org/10.3758/s13428-013-0329-y

Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., Piasecki, T. M., & Watson, D. (2008). Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, *117*(3), 647–661. https://doi.org/10.1037/a0012532

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. https://doi.org/10.1080/10629360600810434

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L. J., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The GRoLTS-checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 451–467. https://doi.org/10.1080/10705511.2016.1247646

van der Maas, H. L. J., Kolstein, R., & van der Pligt, J. (2003). Sudden transitions in attitudes. *Sociological Methods & Research*, *32*(2), 125–152. https://doi.org/10.1177/0049124103253773

van der Maaten, L., Postma, E., & van denHerik, H. (2009). *Dimensionality reduction: A comparative review* (Technical Report TiCC-TR 2009-005). Tilburg University.

van Genugten, C. R., Schuurmans, J., Hoogendoorn, A. W., Araya, R., Andersson, G., Baños, R. M., Berger, T., Botella, C., Cerga Pashoja, A., Cieslak, R., Ebert, D. D., García-Palacios, A., Hazo, J.-B., Herrero, R., Holtzmann,

J., Kemmeren, L., Kleiboer, A., Krieger, T., Rogala, A., … Riper, H. (2022). A data-driven clustering method for discovering profiles in the dynamics of major depressive disorder using a smartphone-based ecological momentary assessment of mood. *Frontiers in Psychiatry*, *13*, 755809. https://doi.org/10.3389/fpsyt.2022.755809

Vasileiadou, E., & Vliegenthart, R. (2014). Studying dynamic social processes with ARIMA modeling. *International Journal of Social Research Methodology*, *17*(6), 693–708. https://doi.org/10.1080/13645579.2013.816257

Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, Montreal, QC, Canada. ACM Press. https://doi.org/10.1145/1553374.1553511

Voelkle, M. C., & Oud, J. H. L. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *The British Journal of Mathematical and Statistical Psychology*, *66*(1), 103–126. https://doi.org/10.1111/j.2044-8317.2012.02043.x

Walls, T. A., Jung, H., & Schwartz, J. E. (2006). Multilevel models for intensive longitudinal data. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 3–37). Oxford University Press. https://doi.org/10.1093/acprof:Oso/9780195173444.003.0001

Wang, L., & Grimm, K. J. (2012). Investigating reliabilities of intraindividual variability indicators. *Multivariate Behavioral Research*, *47*(5), 771–802. https://doi.org/10.1080/00273171.2012.715842

Wang, L., Hamaker, E., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567–581. https://doi.org/10.1037/a0029317

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, *13*(3), 335–364. https://doi.org/10.1007/s10618-005-0039-x

Wardenaar, K. J., & de Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: Towards an empirical solution. *BMC Medicine*, *11*(1), 201. https://doi.org/10.1186/1741-7015-11-201

Weisberg, H. F. (1992). *Central tendency and variability*. Sage Publications.

Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J. J., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, *34*(6), 1060–1072. https://doi.org/10.1002/per.2277

Wenzel, M., & Brose, A. (2023). Addressing measurement issues in affect dynamic research: Modeling emotional inertia's reliability to improve its predictive validity of depressive symptoms. *Emotion (Washington, D.C.)*, *23*(2), 412–424. https://doi.org/10.1037/emo0001108

Wichers, M., Schreuder, M. J., Goekoop, R., & Groen, R. N. (2019). Can we predict the direction of sudden shifts in symptoms? Transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological Medicine*, *49*(3), 380–387. https://doi.org/10.1017/S0033291718002064

Wichers, M., Smit, A. C., & Snippe, E. (2020). Early warning signals based on momentary affect dynamics can expose nearby transitions in depression: A confirmatory single-subject time-series study. *Journal for Person-Oriented Research*, *6*(1), 1–15. https://doi.org/10.17505/jpor.2020.22042

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Appendix A. ESM data challenges and promises

A1. Promises

Time series clustering has a number of conceptual use cases with psychological data. Prime among them is the ability to reduce the time, variable, and person complexity by extracting and organizing participant-level structures. These reduction and structuring qualities can be essential in detecting phenomena and extracting more abstract functional principles (Eronen & Bringmann, 2021). These phenomena and principles can be meaningful differences that distinguish participants in different clusters, as well as important patterns, trends, and relationships that participants share within a cluster (e.g., Schrodt & Gerner, 2000). Once distinct groups and patterns have been identified, researchers can examine the extent to which these within-group and between-group structures are associated with other variables of interest, such as personality traits, demographic characteristics, or other psychological constructs (e.g., Monden et al., 2022). By detecting meaningful and robust structures and patterns, time series clustering can, thus, be used to inform the development of robust theories as well as targeted interventions and therapies for individuals, for example, with mood disorders and other psychological conditions (e.g., Borsboom et al., 2021; Eronen, 2020).

However, while clustering can be incredibly useful, arriving at these clusters critically depends on two core challenges. First, time series need to be made comparable in order to identify key (dis)similarities and second, comparable (dis)similarities need to be accurately distinguishing into different groups (e.g., Aghabozorgi et al., 2015). In practice, most psychological time series cannot be compared based on the raw data itself. This is the case because in most cases the raw time series include too many data points—sometimes referred to as the dimensionality curse (e.g., Altman & Krzywinski, 2018)—and, more importantly, individual time points are oftentimes not directly comparable between participants in psychological data and would lead to misspecifications (e.g., Faloutsos et al., 1994). While

such issues can be avoided with transformations for highly regular, controlled, and comparable time series such as EEG data (e.g., Huang & Jansen, 1985), most ESM researchers are usually not interested in directly comparing individual timepoints between participants but are interested in developmental patterns and relationships.

As a result, most psychological time series are summarized *via* a numerical representation and these numerical summaries are then comparable and used to cluster participants (e.g., Timmerman et al., 2013; see Supplemental Material C). Ideally, the representations that summarize the original time series data should (1) capture the original data accurately without loosing too much information, and (2) should be conceptually meaningful (van der Maaten et al., 2009). Extracting accurate and meaningful representations of the time series can be essential for understanding what goes into the clustering algorithm (i.e., assists with explainability) and can be crucial in making sense of the final cluster output (i.e., assists with interpretability; e.g., Kennedy et al., 2021).

A2. Challenges

We will briefly consider which challenges modern ESM data introduce and what qualities are called for in an extension of the clustering repertoire. We particularly highlight issues of dimensionality, non-equidistant or missing measurements, an interest in non-stationary trends, as well as inconsistent/diverse time scales.

Concerning dimensionality issues, especially more abstract psychological experiences often need a wider variety of measurements to be captured adequately. Today, few clinical conditions are captured with a single symptom measure (e.g., Cramer et al., 2016), emotions are rarely assessed in isolation (e.g., Reitsema et al., 2023), and sociocultural experiences are now widely considered to be multimodal (e.g., Kreienkamp et al., 2024). This also means that modern analysis techniques increasingly need be able to accommodate an increased focus on multivariate developments. At the same time, however, an increase in the number of considered variables tends to come at the expense of computational load for model estimations, and clustering models may not converge (the aforementioned dimensionality curse; Altman & Krzywinski, 2018). A modern time series clustering technique should consequently be able to summarize and structure multivariate phenomena without running into computational load issues.

Another common type of data are measurement regiments that collect data in irregular time intervals (i.e., non-equidistant measurements). Common are, for example, procedures where participants are asked to respond at random times throughout the day (i.e., signal-contingent) or following specific natural events of interest (i.e., event-contingent; see Myin-Germeys et al., 2018; Shiffman et al., 2008). Under such conditions data tends to violate the equidistance assumption that is expected by many time series models (Hamaker & Wichers, 2017). Smaller issues of non-equidistant data can be avoided with transformations (e.g., dynamic time warping, Berndt & Clifford, 1994) or newer modeling procedures (e.g., continuous-time models; de Haan-Rietdijk et al., 2017) but for many analyses, including some cluster approaches, non-equidistant measurements remain a prevalent issue.

Structural missingness remains an even more strenuous challenge. Structural missingness occurs when data is missing because it logically cannot be collected (as opposed to probabilistically missing data; Little & Rubin, 2020; McLean et al., 2017). Often, however, researchers might want to include variables in their models that are not available under all conditions. Follow-up and event-contingent questions are a common example in ESM studies. Researchers, for example, ask about the frequency, intensity, or duration of symptoms—but only if a symptom was present (Kivelä et al., 2022). Such approaches become specifically critical in cases of sensitive questions such as questions about suicidal ideation or other potentially trauma-inducing questions (e.g., Glenn et al., 2022). The most common practice for structurally missing data is to either exclude the variable or any measurement that has no structurally missing data (e.g., Lavori et al., 2008)[8]—neither option suits a research question that wishes to include variables with common structural missingness, such as event-specific or follow-up questions. In short, new clustering approaches should be able to deal with structurally missing data in order to address modern ESM data.

When it comes to studying developmental trajectories, psychological researchers are often also interested in nonstationary processes because they are more representative of the complex, dynamic patterns of the human mind. In psychology, nonstationary processes are typically used to study phenomena such as cognitive development (Quartz & Sejnowski, 1997), decision-making (Ratcliff et al., 2016), and emotion dynamics (Bringmann et al., 2018). These processes are often characterized by changes in the underlying statistical properties of the data over time, such as changes in the mean or variance (Molenaar et al., 2009). Especially when considering changes in mean levels, researchers are often interested in nonlinear changes because they describe human functioning better. For example, in decision making people might switch between choices (Ratcliff et al., 2016), or patients reducing medication might experience mood swings (Helmich et al., 2020). Similarly, psychologists are often also interested in how variances change over time. This is especially the case because several changes in an individual's variance have been found to be indicative of critical changes, including depression relapses and symptom shifts more generally (e.g., Schreuder et al., 2020; Wichers et al., 2020). There is, thus, also a need for time series clustering algorithms that capture nonstationary processes, including nonlinear trends.

Psychological time series often exhibit complex patterns and relationships that can change over different time scales. For example, a time series of daily mood ratings may show a weekly pattern, with higher ratings on the weekends and lower ratings during the week. At the same time, the series may also exhibit a longer-term trend, with overall mood levels increasing or decreasing over the course of several months or years (e.g., Ram et al., 2014). These different time scales can be studied separately or in combination, using different

[8]This is the case because the most commonly used models require complete data (Schafer & Graham, 2002) and structurally missing data cannot be imputed as it logically does not exist (e.g., Lavori et al., 2008).

statistical techniques and modeling approaches (Bertenthal, 2007; Jeronimus, 2019). Different time scales can become an even more difficult issue when different variables in a model develop on different time scales (Bringmann et al., 2022). Different time scales are thus also a concern clustering approaches should be able to address.

It is this background of the common challenges of current ESM data, upon which we propose to consider feature-based clustering. The flexibility of using a wide variety of time series features that represent the important developmental patterns allows users to circumvent many of the issues with multi-dimensionality, non-equidistant or missing measurements, non-stationary trends, as well as diverse time scales.

Appendix B. Validation analyses

To ensure the validity and robustness of our cluster analysis, we conducted a number of additional analyses. In particular, (1) we assessed the impact of our missingness handling, (2) we test a simplified model without dynamic features, and (3) we offer an extended user interface to explore alternative algorithms. The details of these analyses are reported in full detail as part of Supplemental Material A.

B1. Missingness handling

During the variable preparation step, we sought to make the time series comparable and iteratively removed all measurement occasions and participants that had more than 45% missingness. Although this procedure works well for users who wish to use clustering in combination with other parametric models, the 45% threshold might be too conservative if the analysis stands on its own. To test whether this is indeed the case and whether our analysis approach is robust to variations in the missingness handling, we re-ran the main analyses with several more liberal completeness thresholds. We particularly used 0%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, as well as the original 55%; i.e., allowing up to 100% missingness). To compare the results from the different missingness thresholds, we look at the optimal number of clusters k as well as the similarity of the extracted cluster solutions. We find that with almost all missingness thresholds, the optimal number of clusters is 2. The only exceptions are at the thresholds of 0% = 3, 10% = 5, and 25% = 3.

We then compare the clustering results obtained at different thresholds using the Adjusted Rand Index (ARI), which quantifies the similarity between two data clustering assignments. By calculating the ARI for every pair of threshold-based clusterings, we can assess how consistent the cluster assignments are across varying thresholds, even when the number of clusters or their composition changes. This comparison helps us understand the stability of our clustering solution and identify which thresholds yield similar or distinct grouping patterns, providing valuable insight into the robustness of our clustering approach against parameter variations. The ARI is normalized so that −1 indicates perfect disagreement, 0 indicates random (or chance) clustering, and 1 indicates perfect agreement. We find that, except when the optimal solution is larger than two, the cluster similarity is very high (the mean ARI is 0.743 and the mean ARI for all $k = 2$ is 0.892). In short, the number of clusters seems to be a much bigger decision than the cluster assignment itself. Additionally, the high ARI seems to suggest that the PCA and k-means approach for our data is extremely robust to changes in the missingness handling of the raw time series (as long as the number of clusters is the same).

B2. Simplified model

To ensure that the more complex and potentially unreliable dynamic features (e.g., Dejonckheere et al., 2019; Neubauer & Schmiedek, 2020; Wang & Grimm, 2012; Wenzel & Brose, 2023) are necessary to begin with, we additionally check whether a much simpler model with only the central tendency (*median*) and variance (*MAD*) would perform similarly well and would result in a similar separation of the clusters. To compare the model with the main illustration, we assess the performance and similarity of the models. In general, we find that both models perform well across several performance metrics. Additionally, we find a relatively high adjusted Rand index (ARI = 0.758)—indicating that the simplified model separates the two clusters in a similar manner. This similarity is not necessarily surprising given the strong weight of median and MAD in distinguishing the original clusters. Thus, indeed, it is in line with the literature that more complexity is not always necessary (also see Bos et al., 2019).

However, there are two caveats to this preliminary analysis. First, for clinical datasets that look at symptom improvements, a non-stationary trend might be crucial to consider as part of the research question and would probably be present in the data (the same would be true for sudden break-points for episodic conditions). Including more complex dynamic features might thus be crucial for some research questions and will likely depend on the type of ESM data. Second, when we look at the differences in more detail, we see that the original cluster did take the additional features into account in discerning the groups. As an example, the main analysis additionally separates the groups by the linear trend (contrasting an improvement to a deterioration group)—this is less the case for the simpler cluster approach. Arguably, the impact is not as strong for all features and for all variables, but the inclusion of dynamic parameters offers nuanced insights into the temporal patterns and variability not captured by mean levels alone.

B3. Alternative models

Two key decisions during the feature-based clustering approach involve choosing a dimensionality reduction and clustering approach for a given set of data. While a full introduction and evaluation of the many available algorithms is beyond the scope of this paper, we would like to provide some additional insight into the variety of different approaches. To this end, we created an additional resource for readers to interact with the illustration data. As part of this interactive web application, we offer users the option to explore several of the most commonly used dimensionality

and clustering approaches. We have selected four dimensionality reduction algorithms (i.e., PCA, t-SNE, Autoencoders, and UMAP) and three clustering algorithms (i.e., k-means, DBSCAN, and Hierarchical agglomerative clustering). For each of the methods, we have developed an interface that lets users explore the key parameter settings of the algorithms. To provide an introduction to the diversity of possible combinations, we have pre-calculated the performance of the algorithm combinations for common parameter values (showing users a comparison of up to 18,557 model combinations). In the second panel, readers then have the opportunity to explore the cluster results based on their own interaction with the different parameters.

Additionally, we offer users the option to use an unstandardized feature set or the full feature set in cluster analyses. Both of these options are generally not recommended, and the web application aims to give users a more direct understanding of the impact of these decisions. As part of the performance comparison, the application also showcases the sensitivity of different algorithms to highly dimensional data. The DBSCAN algorithm, for example, fails to converge for highly dimensional data (or assigns all points as noise; i.e., the dimensionality curse) and overfits in some parameter ranges. Similarly, hierarchical clustering underperforms with the single linkage method, but only for t-SNE and UMAP dimensionality reduction. The web application thus offers a supplementary resource for readers who wish to explore different analysis approaches within a guided and curated environment. The web application is available as part of our Supplemental Material B or directly at www.tsFeatureClustR.shinyapps.io/webapp/. The full code of the web application is openly available through our GitHub repository (Kreienkamp et al., 2024). Additionally, we offer a further contextualization of the methods and some example guidance on deciding between them in Supplemental Material C.