UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

COMPARING MODEL-BASED, FEATURE-BASED, AND DEEP LEARNING APPROACHES TO TIME-SERIES CLUSTERING FOR INTENSIVE LONGITUDINAL DATA: MONTE CARLO STUDIES AND AN EMPIRICAL DEMONSTRATION

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
MASTER OF SCIENCE

By

THEO MEYER

Norman, Oklahoma
2026

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/343199

COMPARING MODEL-BASED, FEATURE-BASED, AND DEEP LEARNING APPROACHES TO TIME-SERIES CLUSTERING FOR INTENSIVE LONGITUDINAL DATA: MONTE CARLO STUDIES AND AN EMPIRICAL DEMONSTRATION

A THESIS APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Hairong Song, Chair

Dr. Ihnwhi Heo

Dr. Jeongjin Kim

© Copyright by THEO MEYER 2026
All Rights Reserved.

Table of Contents

Abstract viii
1 Introduction 1
2 Literature Review 9
2.1 Representations for Clustering Psychological Time Series 9
2.2 Model-Based Representations 10
2.3 Feature-Based Representations 12
2.4 Deep Learning Representations 14
2.5 Current Comparisons and Their Limits 16
2.6 Cluster-Number Selection and Empirical Validation 17
2.7 The Present Study 19
2.8 What Remains Unanswered 20
3 Research Questions 22
4 Analysis Plan 23
4.1 Simulation Framework and Evaluation Target 23
4.2 Subgroup Recovery 23
4.3 Monte Carlo Summaries and Training-Seed Variation 24
4.4 Clustering Procedures and Performance Summaries 24
4.5 Factor Decomposition and Condition-Level Comparisons 25
4.6 Secondary Diagnostics and Cluster-Number Selection 26
5 Computational Implementation 27
6 Study 1: Common Temporal Order 28
6.1 Method 28
6.2 Results 35
7 Study 2: Mixed Temporal Order 40
7.1 Method 40
7.2 Results 46
8 Robustness Analyses, Extensions, and Diagnostics 49
8.1 Extension: Selection of Temporal Order 49
8.2 Robustness to Subgroup Count 51
8.3 Subgroup-Defining Signal as a Boundary Condition 51
8.4 Reduced-Noise Diagnostic and Latent-Width Robustness Check 52
9 Empirical Demonstration: How Nuts Are the Dutch 53
9.1 Method 53

9.2 Results 57
10 Cross-Study Synthesis 60
11 Discussion 62
11.1 Recovery Levels and Boundary Conditions 63
11.2 Representation Structure and Subgroup Recovery 65
11.3 The Role of Series Length, Separation, and Observation Noise 67
11.4 Fitting Strategy and the Construction of the Representation 69
11.5 When Deep Learning Representations Helped 70
11.6 Subgroup Recovery and Subgroup Enumeration Require Separate Evaluation 71
11.7 Implications for Intensive Longitudinal Research 72
11.8 Limitations 74
11.9 Future Directions 76
12 Conclusion 77
References 79

List of Tables

Table 1. Study 1 Subgroup Recovery by Representation Across 144 Conditions 92
Table 2. Study 1 ANOVA Factor Hierarchy for Model-Based ARI 93
Table 3. Cluster-Number Selection Across 48 Null Conditions and 144 Three-Subgroup Conditions 94
Table 4. Study 2 Subgroup Recovery and Distribution of the Highest ARI Across 162 Conditions 95
Table 5. Study 2 ANOVA Factor Hierarchy for Model-Based ARI (Persistence-Space Design) 96
Table 6. Fitting-Strategy Extension: AR1-Fit ANOVA Factor Hierarchy for Model-Based ARI 97
Table 7. Number of Clusters Selected for Each Representation 98
Table 8. Characteristics of the Two VAR(1) Clusters 99
Table 9. Cluster-Specific VAR(3) Coefficients 100
Table 10. Pre-Diary Measures by Cluster 101

List of Figures

Figure 1. Share of Study 2 Conditions in Which Each Leading Representation Had the Highest ARI Across Series Lengths 102

Abstract

Intensive longitudinal data allow researchers to ask whether individuals form subgroups with different within-person dynamics. This thesis compared five parameter-based, feature-based, and convolutional neural network representations for clustering in two Monte Carlo studies, robustness analyses and extensions, and an empirical demonstration. In order-homogeneous autoregressive populations, direct coefficient estimates produced the strongest subgroup recovery. In mixed-order populations, deep learning representations were more competitive for shorter series, whereas direct estimation gained ground as series lengthened. In Study 1, series length accounted for 50.7% of the variation in Model-Based recovery and sample size for 0.1%. In extensions where scale or volatility defined membership, representations sensitive to those properties performed best. Cluster-number selection was a separate problem: re-expressing the same fitted autoregressive information in redundant, strongly dependent coordinates markedly changed the selected count. In an empirical demonstration adapting the procedure of Li et al. (2025) to affect data, the Bayesian information criterion with a Gaussian mixture selected two clusters for three representations and one for three others, whereas the gap statistic with *k*-means selected one for every representation. Overall, recovery depended on alignment between the representation and the subgroup-defining signal, within-series information, and the geometry presented to the clustering algorithm. Membership recovery and subgroup enumeration should therefore be evaluated separately, as properties of the whole representation-and-algorithm pipeline.

Keywords: intensive longitudinal data, time-series clustering, subgroup recovery, representation learning, vector autoregressive models, cluster-number selection

1 Introduction

Psychological research often characterizes individuals using measurements collected at a single occasion or at a few widely spaced occasions. Such measurements establish differences between people but provide limited information about how an individual's experience unfolds. Intensive longitudinal data make that unfolding process observable. Ecological momentary assessment and experience sampling methodology repeatedly prompt participants to report their current experiences in everyday settings, often several times per day over several weeks (Shiffman et al., 2008). Passive sensing complements these reports with behavioral and physiological records from phones and wearable devices. Together, these approaches support the study of psychological processes in the environments where they occur (Trull & Ebner-Priemer, 2013).

The resulting series distinguish a person's typical level from the dynamics around that level. In affect research, these dynamics include inertia (persistence across occasions), variability (the extent of fluctuation), reactivity to events, and recovery following a disturbance (Hamaker & Wichers, 2017; Kuppens & Verduyn, 2017). Two people with the same average mood may differ in how long distress lasts or how strongly their emotions respond to daily events. These distinctions have psychological relevance: Kuppens et al. (2010), for example, found greater emotional inertia among participants experiencing depression.

Whether such dynamic indices add information beyond mean level is contested. Conventional two-step estimates of person-level inertia can have only moderate reliability, which attenuates their apparent predictive validity. Modeling that unreliability directly recovers part of the association (Wenzel & Brose, 2023). A cross-validated single-cohort study reports a similar limitation: no complex affect-dynamic measure improved out-of-sample prediction of any of 117

psychological and behavioral variables by more than 5.3 percentage points of cross-validated R^2 beyond the mean and standard deviation of positive and negative affect (Shu et al., 2026), extending the pooled finding of Dejonckheere et al. (2019). These results do not identify whether the limitation originates in the reliability of the indices or in the constructs themselves. The present thesis does not assume that dynamic indices are broadly incrementally predictive. Rather, it distinguishes temporal properties from mean level and treats their estimation, use in subgrouping, and validation as empirical questions.

Studying temporal properties requires methods that preserve the ordering of observations and distinguish variation within people from variation between them. Reactivity additionally requires information about events or contexts, because affect reports alone do not identify what produced a change. These requirements link measurement design to the choice of temporal model.

The methodological toolkit begins with models of an individual's series. An autoregressive (AR) model predicts a current observation from earlier observations of the same variable. In a first-order affect model, the lagged coefficient summarizes persistence at the interval being studied. A vector autoregressive (VAR) model extends this approach to several variables, allowing each to depend on its own past and the past of the others. Fitting these models separately for each person allows coefficients to differ across individuals, although estimates from short series can be uncertain. Multilevel AR and VAR models share information across individuals by representing person-specific coefficients through a population distribution. Bringmann et al. (2013) demonstrate this approach for affect dynamics. Random coefficient and multilevel dynamic factor models extend this approach to the dynamics of latent factors measured by several indicators (Song & Ferrer, 2012; Song & Zhang, 2014). Dynamic structural

equation modeling integrates time-series and multilevel structures with structural equation modeling, including latent variables and explicit measurement models (Asparouhov et al., 2018). In conventional random-effects formulations, individuals vary around population relations rather than being assigned discrete memberships.

Network approaches offer a way to interpret these relations as connections among psychological variables. In a temporal VAR network, edges summarize conditional lagged associations and do not by themselves establish causal influence. Networks can be estimated for individuals, populations, or subgroups, so a network perspective does not dictate a single assumption about individual differences. Idiographic approaches emphasize person-specific structure. Group iterative multiple model estimation (GIMME), for example, first identifies relations supported across many individuals, which then guide person-specific model searches (Gates & Molenaar, 2012; Beltz et al., 2016).

Subgrouping methods make shared structure explicit at a level between population-wide and person-specific models. Clusterwise VAR estimates subgroup models together with assignments of individuals to those models (Bulteel et al., 2016). Two-stage approaches estimate individual coefficients first and then cluster the estimates (Ernst et al., 2021). Mixture multilevel VAR permits continuous individual variation within discrete subgroups (Ernst, Timmerman, et al., 2024), while penalized subgrouping distinguishes common, subgroup-specific, and person-specific dynamics (Crawford et al., 2026). For two-stage approaches, a central question is how individual differences in dynamics should be represented: alternative summaries of the same time series may yield different subgroup assignments and selected numbers of clusters.

Interpreting these groupings connects to the longstanding debate over categories and dimensions in personality and psychopathology (Haslam et al., 2012). Intensive longitudinal

research extends that debate to temporal processes. A dimensional account describes continuous differences in persistence, variability, or relations among experiences, as represented through distributions of individual parameters in conventional multilevel models. A categorical account instead describes subgroups with distinct dynamic profiles, evaluated through clusterwise models and clustering procedures. The question is whether a continuous description is sufficient or discrete memberships provide a useful account of individual differences.

Statistical groupings can separate parameter magnitudes without establishing fundamentally different psychological kinds. Their interpretation depends on what the clusters describe and whether those distinctions persist beyond the original analysis. Dimensional and categorical analyses may therefore use the same coefficients while asking different questions: how individuals vary in dynamics, or whether they share distinguishable profiles. This thesis examines the latter while retaining continuous variation as a plausible account when evidence for discrete subgroups is weak.

A hybrid account allows both discrete memberships and continuous differences within subgroups, as in mixture multilevel VAR (Ernst, Timmerman, et al., 2024). That account supplies a further hypothesis to evaluate rather than a conclusion that every population must contain both forms of variation. Liu et al. (2024) sharpen the discussion by distinguishing four assumptions about what individuals share. Strict homogeneity assumes common patterns of relations and common parameter values. Pattern homogeneity allows parameter values to differ while preserving the pattern. Weak homogeneity allows only some processes to generalize across individuals, and no homogeneity assumes no population-level commonalities in those dynamic processes. These assumptions complement the dimensional–categorical distinction: one concerns which relations are shared, and the other concerns how variation across people is organized.

Together, they reframe subgrouping as an issue of generalization scope, rather than a simple choice between estimating one model or many.

The practical consequence is a pooling problem: which observations and individuals should inform a common description? Molenaar (2004) explains why between-person associations cannot automatically be interpreted as within-person processes, identifying restrictive ergodicity conditions for relating population variation to individual temporal variation. A. J. Fisher et al. (2018) provide empirical examples of differences between group-level and person-specific relationships. Neither continuous nor categorical heterogeneity alone determines whether pooling is appropriate. That decision depends on the research target, model specification, and variation accommodated. Multilevel estimation addresses this problem by separating population relations from individual deviations and borrowing information under its assumptions.

Separate models preserve individual variation but can leave substantial estimation uncertainty and require an additional account of how findings generalize. Subgrouping offers an intermediate strategy when common dynamic structure exists within parts of a population. Its value depends on identifying that structure reliably.

Psychological research already uses longitudinal data to identify subgroups, often through differences in trajectories. Group-based trajectory modeling and growth mixture modeling identify latent classes differing in a variable's level and shape over study time (Nagin & Odgers, 2010). These targets differ from grouping by temporal dependence, but the modeling approaches are not mutually exclusive. Ernst, Albers, et al. (2024) place growth curve, growth mixture, and vector-autoregressive models within a framework combining within-person attributes, including growth, cycles, and dynamic interrelations, with continuous and categorical

between-person heterogeneity. The relevant distinction is what defines subgroup membership. In the primary simulations here, within-person standardization removes level and scale as grouping information, leaving temporal dependence as the grouping target.

This choice does not resolve the interpretive risks of extracted classes. Bauer and Curran (2003) showed that growth mixture models could extract additional classes from nonnormal data generated by a single population. Shader and Beauchaine (2022) likewise found in a Monte Carlo evaluation that common class-enumeration statistics frequently failed to recover the generating number of classes under unfavorable separation, sample size, or distributional conditions, and that a single skewed population could be represented by several extracted classes. Multiple components can therefore approximate a distribution without identifying distinct substantive populations. Successful recovery when subgroups exist and restraint when they do not are complementary requirements. Null-based cluster validation addresses the latter (Hennig & Lin, 2015), and this thesis evaluates both for dynamic representations.

Recent psychological research has expanded the approaches to dynamic subgrouping. Li et al. (2025) show that the temporal order specified for person-level VAR models affects membership recovery differently from estimation of subgroup dynamics. Kreienkamp et al. (2025) present feature-based clustering as an accessible framework for comparing psychological series, while Park et al. (2024) examine subgrouping based on lagged and contemporaneous relations. This thesis extends these questions by comparing how selected parameter-based, feature-based, and deep learning representations affect evidence for dynamic subgroup structure.

In the two-stage pipelines evaluated here, each person's series is first represented as a vector, which is then supplied to a clustering algorithm. Model-based representations use fitted parameters, such as autoregressive coefficients. Feature-based representations summarize

properties such as autocorrelations or broader engineered descriptors. Deep learning representations use neural networks to construct coordinates optimized for a training objective. These families differ in the information they preserve, how estimation uncertainty enters their coordinates, and the distances and dependence structure presented to the clustering algorithm. Expanding a scalar estimate into related coordinates can change distances and distributional shape without adding information about that estimate. Comparing representations within and across clustering algorithms reveals whether their relative performance depends on the clustering procedure.

The flexibility of neural representations motivates comparing them with classical approaches, but existing benchmarks do not establish a general advantage. Classification studies identify competitive classical and deep alternatives (Middlehurst et al., 2024), and Paparrizos and Bogireddy (2025) found that no method significantly outperformed k-Shape in their broad clustering evaluation. These comparisons motivate evaluating specific representations against the intended grouping target.

Many clustering benchmarks reuse classification datasets and supply the class count. Their findings do not directly establish how well representations recover subgroups defined by temporal dynamics or select the number of subgroups. The present thesis addresses these questions under common dynamic grouping targets and controlled conditions relevant to psychological data. It also examines how alternative coordinates interact with a fixed cluster-number selection procedure, including transformations that preserve the same estimated information.

Clustering individuals by their dynamics serves three goals: identifying the number of clusters, assigning individuals to clusters, and estimating the dynamics of each cluster. These

goals organize the present thesis. The primary simulations evaluate membership when the number of clusters is supplied, the Unknown-K analysis evaluates selection of the number, and the empirical demonstration shows all three in real data.

The primary objective of this thesis is to determine how time-series representation affects the recovery of subgroups defined by within-person temporal dynamics and the selection of the number of subgroups. Two Monte Carlo studies, robustness analyses and extensions, and an empirical demonstration address these questions. The simulations compare Model-Based, ACF-Empirical, Feature-PCA, CNN-Pred, and CNN-Latent. These five representations use fitted autoregressive coefficients, the autocorrelation function (ACF), principal component analysis (PCA) of engineered features, or a convolutional neural network (CNN), and are evaluated with diagonal-covariance Gaussian-mixture and *k*-means clustering.

Study 1 uses populations with a common temporal order. Study 2 varies order within each subgroup while defining membership in a shared persistence space. Both train CNNs to predict ordinary least squares (OLS) coefficient estimates from independently generated series rather than subgroup labels. The studies also differ in population structure, fitting specification, engineered features, and series-length range, so their comparison does not isolate order heterogeneity.

The separate Unknown-K analysis shares Study 1's factor structure and evaluates cluster-number selection, including how different representations of the same fitted information affect that selection. Robustness analyses and extensions vary the number of generating subgroups, the fitting strategy applied before clustering, and the property that defines membership, including scale- and volatility-defined subgroups. The empirical demonstration adapts the procedure of Li et al. (2025) to How Nuts Are the Dutch experience-sampling data, adding the five

representations to Li et al.'s VAR(1) and VAR(3) representations of positive and negative affect. It reports the number of clusters selected by two selection procedures, the agreement between the resulting memberships, and the dynamics and pre-diary characteristics of the clusters. The simulations evaluate recovery against known memberships, whereas the demonstration shows how the pipelines behave in real data, where memberships are unknown.

2 Literature Review

2.1 Representations for Clustering Psychological Time Series

Time-series clustering requires a definition of similarity suited to the research question, whether it concerns trajectories, temporal dependence, distributional properties, or their combination. Classical surveys distinguish approaches using raw series, extracted features, or fitted models (Liao, 2005; Aghabozorgi et al., 2015). These categories overlap: a fitted coefficient is also a feature, and a network can learn to estimate it. The useful distinction is therefore how a representation is constructed and what information it preserves (Van Mechelen et al., 2024; Paparrizos et al., 2024). Representation choice is especially consequential for short, noisy, heterogeneous psychological series, whose clustering must be evaluated against a scientific goal rather than a single notion of true clusters (Hennig, 2015).

This thesis evaluates two-stage pipelines that represent each series as a vector before clustering. Methods that jointly estimate temporal models and memberships are reviewed as context, because they use information differently. For each representation family, the review explains its rationale and limitations relevant to the simulation design.

2.2 Model-Based Representations

Model-based approaches represent each series through fitted time-series parameters. In the low-order autoregressive settings examined here, coefficients compactly summarize the specified dynamics. Piccolo (1990) compared invertible autoregressive integrated moving average processes using distances between their autoregressive expansions. Maharaj (2000) treated similarity as an inferential question, testing whether two series plausibly share a generating model. Xiong and Yeung (2004) fitted mixtures of autoregressive moving-average models using expectation–maximization. Ren and Barnett (2022) instead cluster through a probabilistic autoregressive mixture over autocovariance structures, rather than separately clustering individual coefficient estimates. These methods share a modeling rationale but differ in implementation.

Psychological applications of VAR models include several approaches to subgrouping. Bulteel et al. (2016) introduced clusterwise VAR modeling, which jointly estimates subgroup dynamics and memberships. Ernst et al. (2021) compared a two-stage approach that clusters individually estimated VAR coefficients using Gaussian mixtures with an alternating-least-squares procedure. Their comparison included simulations and Dutch experience-sampling data, where the number of clusters was selected empirically. Ernst, Timmerman, et al. (2024) developed mixture multilevel VAR modeling to accommodate discrete subgroups and continuous individual variation within them. Li et al. (2025) compared higher- and lower-order representations under Gaussian-mixture and *k*-means clustering. In their simulations, lower order favored membership recovery, whereas higher order favored estimation of subgroup dynamics. Park et al. (2024) examined chain graphical VAR models incorporating lagged and contemporaneous relations. Crawford et al. (2026) extended the penalized multi-VAR

framework of Z. F. Fisher et al. (2022) to common, subgroup-specific, and person-specific dynamics. These developments illustrate ongoing model-based subgrouping research. The present thesis compares selected two-stage representations rather than benchmarking joint and penalized procedures.

These methods differ along two dimensions: whether parameters and memberships are estimated jointly or sequentially, and whether individuals within a subgroup share parameters or vary around subgroup-level relations. These distinctions cross, because joint models can allow individual variation and two-stage methods can represent differences among individual estimates. Clusterwise modeling, multilevel mixtures, and two-stage clustering therefore concern different aspects of the analysis. Joint estimation can share information across people, but parameter and membership uncertainty still depend on the number of observations. Two-stage representations inherit estimation uncertainty directly: autoregressive estimates have finite-sample bias and sampling variability that decrease with series length (Cheang & Reinsel, 2000; Kilian, 1998; Shaman & Stine, 1988). Short series can therefore obscure subgroup separation before clustering begins.

Coefficient representations align directly with the grouping target when subgroups differ in autoregressive dynamics. Their limitations concern estimation precision and the properties they represent. Short series yield uncertain coefficients, while coefficient-only vectors do not explicitly represent innovation variance, distributional shape, or other omitted properties. Such properties may nevertheless affect the estimates' sampling behavior and influence the resulting partition. Longer series reduce sampling uncertainty but do not resolve misspecification. Performance therefore depends on whether the selected coefficients preserve the subgroup-defining distinctions accurately enough for the downstream clustering algorithm.

2.3 Feature-Based Representations

Feature-based approaches summarize descriptive properties without committing to a single model. X. Wang et al. (2006) developed characteristic-based clustering using measures of trend, seasonality, serial correlation, skewness, and nonlinearity. Later resources broadened this approach: hctsa extracts thousands of features across methodological traditions (Fulcher et al., 2013; Fulcher & Jones, 2017), catch22 provides a compact set selected for informativeness and minimal redundancy (Lubba et al., 2019), and tsfresh automates extraction with relevance assessment (Christ et al., 2018). Feature spaces also characterize conditions favoring different forecasting methods (Kang et al., 2017) and, after principal component reduction, identify unusual series in large collections (Hyndman et al., 2015). These resources illustrate the family's scope but are not the pipelines evaluated here.

Kreienkamp et al. (2025) provide a practical framework for feature-based clustering of psychological time series. Features specify the properties used to compare participants, but their interpretability depends on selection and subsequent transformation. Dimension reduction can obscure otherwise interpretable summaries, while a small feature set may omit relevant distinctions. Feature selection is therefore a substantive modeling decision rather than merely a preprocessing step.

Autocorrelation representations summarize associations at selected lags and are established tools for time-series clustering (Galeano & Peña, 2000; Montero & Vilar, 2014). Empirical ACF features are sample autocorrelations computed from observed series. Model-implied ACF features are derived from fitted parameters: for a stationary AR(1) process, $\rho_k = \varphi^k$, giving the estimated vector $(\hat{\varphi}, \hat{\varphi}^2, \ldots, \hat{\varphi}^{10})$. Sample autocorrelations need not satisfy this relation and are not deterministic transformations of the fitted coefficient.

This distinction separates comparisons of estimators from comparisons of representations of fixed estimated information. Empirical autocorrelations are multiple statistics with a joint sampling distribution. The implied vector retains the fitted estimate in its first coordinate and adds deterministic functions of it, contributing no further information about that estimate. However, the transformation changes distances and distributions and need not preserve the clustering algorithm's assumptions or resulting partition. Both approaches require choosing a lag window. In short series, unstable higher-lag estimates can add noise faster than signal.

Feature count can exceed the amount of distinct information represented. Henderson and Fulcher (2021) found substantial redundancy within and overlap between seven feature sets. Across fifteen experience-sampling studies, Dejonckheere et al. (2019) found that common affect-dynamic measures were highly interdependent and added little to prediction of the examined well-being outcomes beyond affect means and variances.

Predicting well-being and recovering dynamic subgroups are different objectives, so these findings do not establish that feature representations are unsuitable for clustering. They motivate examining distinct information, coordinate dependence, and conditioning in feature spaces. Their implications for cluster-number selection are considered below. Dimension reduction can address redundancy, but principal components preserve directions of variance that need not distinguish subgroup memberships.

The simulations evaluate two feature representations: ten empirical autocorrelations and study-specific engineered features transformed by principal components analysis. The comparisons therefore concern these implementations and their scaling and reduction steps, not catch22, hctsa, or the workflow of Kreienkamp et al. (2025). Moreover, several engineered

features come from fitted models, so “feature-based” describes their construction without implying an absence of model assumptions.

2.4 Deep Learning Representations

Deep learning representations use neural networks to map time series into coordinates optimized for a training objective. Their flexibility is constrained by the architecture, input processing, objective, and training population. Consequently, a representation useful for one task may preserve little information relevant to another. Convolutional networks detect local temporal patterns and combine them into higher-level features, providing strong time-series classification baselines (Z. Wang et al., 2017). Ismail Fawaz et al. (2019) review their architectures and training regimes, while much benchmarking uses the University of California, Riverside (UCR) archive (Dau et al., 2019). For psychological subgrouping, the question is whether learned coordinates preserve relevant between-person differences in dynamics. Three approaches are considered here.

The first approach trains networks on estimated coefficient targets. Both simulation studies use convolutional networks trained on OLS estimates from independently generated series. Each network supplies predicted coefficients, CNN-Pred, and an internal embedding, CNN-Latent. These are two representations from the same network within each study.

Supervision targets coefficient estimates rather than subgroup memberships, so the learned mapping may reproduce, contract, or otherwise transform the estimator’s properties. The loss directly constrains predicted coefficients. In contrast, the embedding supports those predictions but may retain additional variation and lacks a coordinate-wise parameter interpretation. Its distances therefore need not correspond to distances between generating parameters. Evaluating both representations distinguishes the usefulness of explicit predictions

from that of internal coordinates. Cross-study differences reflect changes in data generation and representation construction rather than supervision type.

The second approach uses self-supervised objectives derived from the series. Franceschi et al. (2019) combine a causal dilated convolutional encoder with a triplet objective, Zerveas et al. (2021) use masked reconstruction in a transformer, and Yue et al. (2022) develop hierarchical contrastive learning in TS2Vec using augmented views. Reconstruction, contrastive discrimination, and forecasting reward different relationships among observations. Unlike fitted coefficients, these representations do not directly parameterize dynamics, and whether they preserve between-person differences in persistence requires empirical evaluation.

The third approach, deep clustering, jointly learns representations and partitions. Examples combine an autoencoder with a clustering loss (Madiraju et al., 2018), and Lafabregue et al. (2022) comparatively review the broader family. Large-scale pretraining is a separate design choice rather than a further kind of objective: foundation models such as Chronos (Ansari et al., 2024) and MOMENT (Goswami et al., 2024) are pretrained on separate collections of series and applied or adapted downstream. These approaches differ in whether the target sample contributes to training and whether learning explicitly optimizes a partition. The present thesis evaluates coefficient-targeted representations, with self-supervised learning, deep clustering, and foundation models providing context.

Deep learning representations offer flexibility by estimating features within a chosen architecture and objective rather than specifying each feature individually. However, the objective may not align with the subgroup-defining property, and training typically requires more series than a psychological study provides. Representation dimensionality and geometry also depend on the architecture, objective, and training distribution, not solely on the generating

process. Thus, learning useful features does not guarantee that their coordinate structure suits the downstream clustering procedure.

2.5 Current Comparisons and Their Limits

Comparisons among time-series methods must be interpreted relative to the task being evaluated. Forecasting concerns prediction of future observations, classification evaluates recovery of supplied labels, and clustering seeks a partition under a selected notion of similarity. A representation that performs well for one objective therefore need not provide a useful geometry for another. The point matters here because the target is not generic time-series similarity but recovery of subgroups defined by latent temporal dynamics.

Large classification benchmarks show that both classical and deep learning representations can be competitive and provide little support for universal deep-learning superiority (Bagnall et al., 2017; Middlehurst et al., 2024). Clustering comparisons point in the same direction. Lafabregue et al. (2022) reviewed deep representation-learning approaches across architectures and objectives, and Paparrizos and Bogireddy (2025) compared classical, deep-learning, and foundation-model approaches across a large collection of datasets, finding no general dominance of the learned alternatives. These studies establish the importance of comparing concrete implementations under common conditions, but their external class labels and supplied cluster counts do not determine which representation best recovers subgroups defined by autoregressive dynamics.

Psychological research addresses a problem closer to the present one. Ernst et al. (2021) compared model-based subgrouping strategies and examined cluster-number decisions in empirical emotion data. Li et al. (2025) showed that temporal-order specification can affect membership recovery differently from estimation of subgroup dynamics and applied their

methods to experience-sampling affect data. Related work has examined feature-based psychological time-series clustering and heterogeneous multivariate dynamics (Kreienkamp et al., 2025; Park et al., 2024; Crawford et al., 2026). Collectively, this literature demonstrates that representation choice, model specification, and subgroup enumeration are consequential, but it does not provide a common comparison of the parameter-based, engineered, and CNN-derived representations evaluated here under the same dynamic grouping targets.

The present studies provide that comparison. Series length varies the within-person information available for estimating dynamics, sample size the number of individuals available for clustering, subgroup separation the distinctiveness of the generating subgroups, and observation noise the clarity with which the latent process is observed. Study 1 evaluates these representations when temporal order is common within a dataset. Study 2 allows order to vary within every subgroup while defining membership in a shared persistence space, which separates recovery of the intended subgroup structure from recovery of temporal order. Because the studies also differ in population structure, autoregressive fitting specification, Feature-PCA construction, CNN training-sample size, and the series-length range evaluated, cross-study differences cannot be attributed to mixed temporal order alone.

2.6 Cluster-Number Selection and Empirical Validation

Recovering memberships with a supplied cluster count differs from deciding whether multiple clusters are warranted. Fixed-count recovery can be scored against known memberships but cannot assess selection of additional clusters under a no-subgroup population. Hennig and Lin (2015) use a parametric-bootstrap framework to assess homogeneity and cluster number, emphasizing that the null must specify which population structure is retained.

Two controls are used here. A constant-parameter population assigns identical dynamics to everyone, whereas a continuously heterogeneous population allows variation without discrete subgroups. These controls address different forms of no-subgroup structure. A Gaussian mixture may approximate a continuous non-Gaussian distribution using several components. Consequently, a selected component count does not establish the number of psychological types, and selecting one component does not prove identical individual dynamics.

Representation structure requires distinguishing four concepts. Ambient dimension counts the coordinates supplied to the clustering algorithm, whereas intrinsic dimension describes the dimension of the structure supporting the observations. Covariance rank concerns linearly independent covariance directions, while effective rank summarizes how variance is distributed across those directions, for example by the participation ratio. A curved one-dimensional structure can have full covariance rank. Coordinate dependence can therefore matter for diagonal-covariance Gaussian mixtures even without exact rank deficiency. The Unknown-K analysis, which used full-covariance mixtures, examines how one such transformation interacts with Gaussian-mixture selection by the Bayesian information criterion (BIC), without assuming that dimension alone determines the selected component count.

Deterministic transformations such as model-implied ACF change the coordinates of an original estimate without adding estimated dynamic information. Comparing such a transformation with the original estimate therefore examines representation structure while holding the estimated information fixed.

Empirical validation must then address what can be learned without generating memberships. Agreement across representations measures consistency between analyses, not accuracy against a partition. Replication across disjoint temporal blocks assesses whether a

grouping persists across portions of the observation period, whereas agreement between nested windows partly reflects their shared observations. External associations address links with other measured characteristics. Ullmann et al. (2022) systematize these validation aims. Each provides complementary evidence, and failure on one can reflect limited information as readily as an unsuitable grouping.

Positive and null controls help interpret such validation results. Positive controls indicate which separations the implemented pipeline can detect under the evaluated conditions, whereas null controls characterize apparent structure in the absence of target subgroups. In this thesis, both kinds of condition are part of the simulated Unknown-K analysis; the empirical demonstration does not include them. Irregular observation intervals, missing reports, and imputation also determine which temporal information enters an empirical representation. Such a representation must therefore be distinguished from complete simulated observations, and preprocessing and eligibility decisions are part of the interpretation. This framework allows both positive and limited evidence to be reported without treating either a partition or a one-cluster solution as a definitive description of the population.

2.7 The Present Study

The two main simulations compare five representations drawn from four pipelines. Model-Based uses fitted autoregressive coefficients. ACF-Empirical uses sample autocorrelations at ten lags. Feature-PCA uses study-specific engineered feature sets transformed using principal components analysis, with two components retained in Study 1 and ten in Study 2. CNN-Pred uses the predicted coefficients of a coefficient-targeted convolutional network, and CNN-Latent uses that network's internal embedding, so the CNN pipeline supplies two representations. Each is evaluated with diagonal-covariance Gaussian-mixture and *k*-means

clustering, so comparisons are made within each downstream procedure. The networks use OLS-estimated coefficient targets obtained from independently generated training series and are not trained as classifiers of the generating subgroup labels.

Study 1 provides a baseline in which individuals within a dataset share a temporal order, while coefficients vary across subgroups and individuals. Study 2 defines subgroups in a shared persistence space and includes both temporal orders within each subgroup. Order is therefore not the subgroup label. The empirical demonstration applies the five representations, together with the VAR(1) and VAR(3) representations of Li et al. (2025), to positive- and negative-affect reports from the How Nuts Are the Dutch project.

2.8 What Remains Unanswered

The literature motivates five questions at the intersection of representation construction, the information available in individual series, and subgroup validation. The contribution of the thesis is threefold. First, model-based, feature-based, and deep learning representations are compared side by side on the same subgroups. Second, the comparison uses conditions common in psychological data: short, noisy series and, in Study 2, mixed temporal order within subgroups. Third, an empirical demonstration adds these representations to the procedure of Li et al. (2025) and reports the number of clusters, the memberships, and the cluster-level dynamics in real data. In this way, the thesis builds on prior work comparing methods and examining empirical clusters.

Gap 1: Comparison under a shared dynamic grouping target. Extensive methodological alternatives exist, but directly comparable evidence remains limited for the selected parameter-based, feature-based, and CNN-derived representations under common autoregressive grouping conditions. A shared evaluation within each clustering algorithm is

needed to assess their dependence on series length, sample size, separation, and observation noise.

Gap 2: Mixed temporal order within subgroups. Temporal-order selection is an established concern, but comparative evidence remains limited when order varies within each subgroup defined by a shared persistence target. This design separates recovery of the intended subgroups from identification of temporal order.

Gap 3: Recovery and cluster-number calibration. Performance with a supplied cluster count does not resolve performance when the count must be selected. The representations require joint evaluation of recovery under a specified number of clusters and of selection behavior under the homogeneous or continuously heterogeneous controls appropriate to each analysis.

Gap 4: Representation structure under a fixed selection procedure. Additional coordinates can introduce information, repeat it, or transform its distribution. Controlled comparisons are needed to distinguish these possibilities and to examine how redundancy and covariance structure interact with Gaussian-mixture selection by the Bayesian information criterion.

Gap 5: Representations in empirical data without known memberships. Existing experience-sampling applications establish the relevance of subgrouping. Li et al. (2025) applied VAR-based clustering to affect data from the How Nuts Are the Dutch project, but their application did not include the parameter-based, feature-based, and CNN-derived representations evaluated here. Adding them to the same procedure shows whether the selected number of clusters, the resulting memberships, and the cluster-level dynamics depend on the representation and the selection procedure in real data.

3 Research Questions

RQ1 and RQ2 are the primary questions and are addressed by Studies 1 and 2. RQ3 and RQ4 concern selection of the number of subgroups and are addressed mainly by the Unknown-K analysis, and RQ5 is addressed by the empirical demonstration.

RQ1. Within common temporal-order conditions with known subgroup memberships, how accurately do the five representations recover membership, and how does their relative performance depend on series length, sample size, subgroup separation, observation-noise level, and the downstream clustering algorithm?

RQ2. How does the relative performance of the representations change when subgroups are defined in a shared persistence space and each subgroup contains individuals with different temporal orders?

RQ3. When the number of subgroups must be selected, how do the representations perform in recovering the generating count and in avoiding additional clusters under the no-subgroup controls specified for each analysis?

RQ4. When estimated dynamic information is held fixed, how do changes in its representation, including redundant coordinates and their dependence structure, affect cluster-number selection by the Bayesian information criterion for Gaussian mixtures?

RQ5. In empirical experience-sampling affect data without known memberships, how do the selected number of clusters, the resulting memberships, and the cluster-level dynamics depend on the representation and the selection procedure when the five representations are added to the procedure of Li et al. (2025)?

4 Analysis Plan

4.1 Simulation Framework and Evaluation Target

Both studies used Monte Carlo simulation to evaluate subgroup recovery under known data-generating conditions. Simulation provided generating memberships against which recovered partitions could be compared while systematically varying properties of the data and clustering problem. The studies followed the simulation framework of Morris et al. (2019), covering aims, data-generating mechanisms, estimands, methods, and performance measures.

4.2 Subgroup Recovery

The primary performance measure was the adjusted Rand index (ARI; Hubert & Arabie, 1985), which quantifies agreement between two partitions while adjusting for chance agreement under its reference model. In the simulations, ARI compared recovered memberships with the generating subgroup labels. ARI equals 1 for identical partitions, is near zero when agreement approaches its reference expectation, and can be negative when agreement is below that expectation. Because ARI is invariant to permutations of cluster labels, label alignment was unnecessary for the primary recovery analyses. When cluster-specific diagnostics required correspondence between recovered clusters and generating subgroups, labels were aligned using the Hungarian algorithm (Kuhn, 1955).

Interpretation emphasized the magnitude and pattern of ARI rather than categorical thresholds. Small differences among methods near the floor were distinguished from differences producing substantively greater recovery, and no published ARI value was treated as a universal criterion for adequate clustering (Milligan & Cooper, 1986; Steinley, 2004). The Discussion uses Steinley's (2004) recovery bands only as descriptive reference points.

4.3 Monte Carlo Summaries and Training-Seed Variation

For each condition, representation, and clustering algorithm, performance was summarized using the mean and standard deviation of replication-level ARI. The Monte Carlo standard error (MCSE) of each condition mean was calculated as

$$\text{MCSE}(\text{mean ARI}) = s(\text{ARI})/\sqrt{R},$$

where s(ARI) is the standard deviation of replication-level scores and R is the number of replications. MCSE therefore quantifies uncertainty arising from the finite number of Monte Carlo replications rather than variation across experimental conditions (Morris et al., 2019).

CNN-derived representations were evaluated using five independently trained networks per condition. Each network generated CNN-Pred and CNN-Latent representations for every evaluation dataset. These were clustered and scored separately, and the five resulting ARI values were averaged within each replication before condition-level aggregation. Performance scores, rather than latent coordinates, were averaged because independently trained latent spaces need not share aligned coordinates. This procedure incorporated performance across networks without treating repeated evaluations of the same dataset as independent replications.

4.4 Clustering Procedures and Performance Summaries

Each primary representation was evaluated using two downstream clustering procedures: a diagonal-covariance Gaussian mixture model (GMM) and *k*-means. Both were supplied with $K = 3$ for the primary recovery analyses. The diagonal GMM estimated coordinate-specific component variances but no within-component covariance terms, making its performance potentially sensitive to dependence and orientation within the representation. The *k*-means algorithm instead formed partitions by minimizing within-cluster squared Euclidean distances. Evaluating both procedures allowed representation effects to be examined within a common

clustering algorithm while also testing whether relative performance depended on the downstream clustering procedure.

Two complementary summaries described representation performance. First, pooled mean ARI summarized average performance across each study's conditions. When a single clustering algorithm is reported for a representation, it is the pooled-best clustering algorithm, defined by the higher pooled mean ARI across conditions. Results for the GMM and *k*-means were also reported separately.

Second, condition-level counts described how often each representation achieved the highest available mean ARI. Each representation used its within-condition best clustering algorithm, defined as the clustering algorithm with the higher mean ARI for that representation and condition, before the representation with the highest mean ARI was identified. Thus, pooled means used one clustering algorithm per representation across a study, whereas the condition-level counts allowed that choice to vary by condition. The two rules are distinguished when reporting these counts. Both summaries are retrospective evaluations based on known memberships, not empirical algorithm-selection procedures.

4.5 Factor Decomposition and Condition-Level Comparisons

Descriptive analysis of variance (ANOVA) models summarized how condition-level mean ARI varied across the manipulated factors. Interpretation focused on η^2, expressed as each term's sum of squares relative to the total used in the decomposition, rather than on conventional significance testing (Langsrud, 2003; Richardson, 2011). Main effects were interpreted in conjunction with interactions when the latter were substantively important.

Because Study 1 used 1,000 replications per condition and Study 2 used 500, replication-level tests could detect differences too small to be substantively meaningful. Inferential

comparisons therefore used simulation conditions as the unit of analysis. Within each study, a Friedman test (Friedman, 1937) compared the five representations, followed by two-sided Wilcoxon signed-rank comparisons (Wilcoxon, 1945) of Model-Based with each competitor. The four comparisons were Holm-corrected (Holm, 1979) within each study, with matched-pairs rank-biserial correlations reported as effect sizes. These analyses assessed whether differences were consistent across conditions rather than concentrated in a few.

The inferential comparisons used each representation's pooled-best clustering algorithm, as did the pooled means. Condition-level counts instead allowed the better clustering algorithm to vary by condition. These aggregation rules are not interchangeable.

4.6 Secondary Diagnostics and Cluster-Number Selection

Secondary diagnostics examined parameter prediction, mixture-model convergence, and characteristics of recovered memberships to distinguish retention of relevant temporal information from successful subgroup recovery.

The primary recovery analyses fixed $K = 3$ to isolate membership recovery from cluster-number estimation. The Unknown-K analysis addressed the separate problem of selecting the number of components, evaluating both recovery of the generating count when discrete subgroups were present and additional-component selection under no-subgroup conditions. Forced $K = 3$ null conditions were not interpreted as false-positive cluster-selection tests because any nontrivial forced partition of a single generating class yields ARI of zero under the scoring setup. Conclusions about selection of additional clusters therefore came from analyses in which the number of components was allowed to vary.

5 Computational Implementation

The analyses were implemented in Python. NumPy (Harris et al., 2020) provided numerical operations, scikit-learn (Pedregosa et al., 2011) provided preprocessing and clustering, and PyTorch (Paszke et al., 2019) provided neural-network training and evaluation. Autoregressive estimation used statsmodels (Seabold & Perktold, 2010) for Study 1 Model-Based fits, a separate vectorized least-squares routine for Study 1 CNN targets, and NumPy least squares for the common AR(2) fits in Study 2. Neural computations used available accelerator hardware, with CPU execution when no accelerator was available.

Training data, diagnostic data, and Monte Carlo evaluation data were generated using distinct seed assignments. The separation between these datasets was maintained throughout the primary comparisons. Networks were trained on independently generated series rather than on the datasets used to calculate subgroup recovery.

In the primary analyses, the five networks trained for each condition were reused across that condition's evaluation replications, keeping network weights fixed while the evaluation datasets changed.

Both primary simulation studies trained their CNNs on OLS coefficient estimates obtained from independently generated training series. Study 1 used coefficient targets corresponding to the temporal order of the condition, whereas Study 2 used estimates from the common AR(2) specification employed for its mixed-order design. The true generating coefficients available in simulation were retained for diagnostic evaluation but were not used as CNN training targets. Thus, the neural supervision incorporated properties of OLS estimation in both primary studies.

For AR(2) generation, coefficient draws were checked against stationarity constraints, and invalid draws were rejected and resampled. The two studies handled exhaustion of that loop differently. Study 1 assigned the subgroup centroid to any individual for whom no admissible pair was obtained within 1,000 attempts, whereas Study 2 raised an error rather than substituting a value. AR(1) coefficients were not rejected. Each drawn coefficient was instead clipped to the interval [−0.99, 0.99] in both studies. The accepted coefficient distributions therefore reflected these restrictions in addition to the initial perturbation distributions.

Each network was trained for 100 epochs using the Adam optimizer (learning rate .001), mean squared error loss, Xavier initialization, mini-batches of 32, and a reduce-on-plateau learning-rate schedule (factor 0.5, patience 15 epochs) driven by mean training loss. No early stopping or validation split was used, so every network completed the full training budget.

Held-out coefficient-prediction diagnostics used 10,002 independently generated series per condition. Subgroup-recovery analyses used the condition-specific sample sizes and replication counts. Gaussian mixtures used 10 initializations and covariance regularization of 1e-4, while *k*-means also used 10 initializations. Within each replication, all representations were evaluated on the same dataset, making comparisons paired. Seeds for the clustering algorithms were reproducible, although identical seed values did not imply identical initializations across different feature spaces.

6 Study 1: Common Temporal Order

6.1 Method

Experimental Design. Study 1 evaluated subgroup recovery in order-homogeneous populations using a fully crossed Monte Carlo design. Within each simulated dataset, all

individuals followed the same autoregressive family, either AR(1) or AR(2). This provided a baseline before introducing mixed-order populations in Study 2.

The design crossed sample size ($N \in \{100, 300\}$), series length ($T \in \{15, 50, 100, 200\}$; Hamaker & Wichers, 2017), temporal order {AR(1), AR(2)}, and signal-to-noise ratio (SNR $\in \{1, 2, 4\}$). Subgroup separation varied across Easy, Medium, and Hard levels. A clean-baseline diagnostic ($N = 300$, $T = 100$, AR(1), Easy separation) contained neither within-subgroup coefficient variation nor measurement noise. It provided a recovery benchmark under favorable conditions and was excluded from pooled summaries.

No-subgroup conditions were excluded from this fixed-K design because forcing $K = 3$ on a single generating class produces an ARI of zero by construction and cannot test false-positive subgroup detection. The separate Unknown-K analysis evaluated null behavior with a variable component count. Study 1 therefore contained 145 conditions: 144 $K = 3$ clustering conditions and one clean-baseline diagnostic. Each condition was replicated 1,000 times.

Conditions and Data Generation. Latent time series were generated from stationary AR(1) and AR(2) processes (Brockwell & Davis, 1991). The AR(1) process was defined as

$$z_t = \varphi\, z_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim N(0, \sigma^2_\varepsilon),$$

and the AR(2) process was defined as

$$z_t = \varphi_1\, z_{t-1} + \varphi_2\, z_{t-2} + \varepsilon_t.$$

For AR(2) conditions, stationarity was enforced by requiring each sampled parameter pair to satisfy the standard triangle constraints,

$$\varphi_2 + \varphi_1 < 1, \;\; \varphi_2 - \varphi_1 < 1, \;\; |\varphi_2| < 1.$$

Invalid AR(2) parameter pairs were rejected and redrawn. The fallback and AR(1) clipping procedures are described in Section 5. For the clustering conditions, person-specific

autoregressive parameters were sampled around subgroup centroids with a perturbation standard deviation of $\tau = 0.10$. Because clipping and stationarity restrictions affected the accepted coefficients, τ describes the perturbation distribution before those restrictions rather than the realized within-subgroup dispersion. Cluster sizes were balanced as closely as possible, with exact balance at $N = 300$ and near-balance at $N = 100$.

For AR(1), subgroup separation was operationalized through a separation-to-spread ratio defined as

$$SSR = |\mu_{k+1} - \mu_k| / (2\tau),$$

where μ_k is the cluster centroid and $\tau = 0.10$ is the perturbation standard deviation. Easy, Medium, and Hard separation corresponded to SSR values of 2.0, 1.5, and 1.0, with AR(1) centroid sets {0.1, 0.5, 0.9}, {0.2, 0.5, 0.8}, and {0.3, 0.5, 0.7}, respectively. For AR(2), separation was defined analogously in two-dimensional parameter space using a bivariate separation-to-spread ratio,

$$SSR_{2D} = min_{j \neq k} \|\mu_j - \mu_k\| / (2\tau\sqrt{2}),$$

where $\|\mu_j - \mu_k\|$ is the Euclidean distance between the nearest pair of centroids in (φ_1, φ_2) space and the denominator uses the perturbation standard deviation scaled for two dimensions. Easy, Medium, and Hard AR(2) centroid triplets were {(0.0, 0.0), (0.5, 0.2), (0.8, −0.3)}, {(0.2, 0.1), (0.5, 0.3), (0.2, 0.5)}, and {(0.3, 0.2), (0.5, 0.3), (0.25, 0.5)}, respectively. Under this definition, the Easy, Medium, and Hard AR(2) triplets give SSR_{2D} values of approximately 1.9, 1.3, and 0.8. These levels are author-chosen design values, selected to span a practical range from clearly recoverable subgroup structure to substantial within-population overlap. Prior work motivates manipulating separation and sample size in cluster-recovery simulations (Steinley, 2004; Dalmaijer et al., 2022) but does not establish these particular values.

The no-subgroup conditions of the separate Unknown-K analysis used the same sample-size, series-length, temporal-order, and SNR levels but had no subgroup-separation factor. They were generated from a one-component solution centered at $\varphi = 0.50$ for AR(1) and $(\varphi_1, \varphi_2) = (0.50, 0.20)$ for AR(2).

Signal-to-noise ratio was imposed by combining the z-normalized latent series with additive Gaussian measurement error. The observed series was modeled as

$$y_t = \alpha\, z_t + e_t,$$

where $\alpha = \sqrt{(\mathrm{SNR} / (1 + \mathrm{SNR}))}$ and $e_t \sim N(0, 1 / (1 + \mathrm{SNR}))$. The signal fraction was SNR / (1 + SNR), the noise fraction was 1 / (1 + SNR), and the observed series had unit variance by construction up to finite-sample drift. SNR levels of 1, 2, and 4 corresponded to measurement-error variances of 0.50, 0.33, and 0.20, respectively. A burn-in period of 200 observations was discarded before retaining the series of length *T*. Each retained series was then z-normalized before feature extraction so that clustering emphasized temporal structure rather than raw mean or variance differences. The design allowed recovery to be examined across series information, subgroup separation, measurement quality, and autoregressive complexity.

Representations Compared. Study 1 compared five representations. Each was constructed from the same observed series within a replication and evaluated using both clustering procedures.

Model-Based. The Model-Based representation consisted of OLS estimates from the autoregressive model matching the generating order. In AR(1) conditions, each individual was represented by

$$x_i = (\hat{\varphi}_i).$$

In AR(2) conditions, the representation was

$$x_i = (\hat{\varphi}_{1i}, \hat{\varphi}_{2i}).$$

Model-Based therefore used one or two coordinates corresponding directly to the parameters that define the subgroups, subject to finite-series bias and sampling variability in their estimation (Shaman & Stine, 1988). These estimates were obtained from autoregressive models fitted without an intercept to the standardized series. In AR(1) conditions the fitted coefficient was clipped to the interval [−0.99, 0.99] before clustering, whereas in AR(2) conditions both fitted coefficients were retained without componentwise clipping, because a first coefficient greater than one can be admissible within the stationarity triangle.

ACF-Empirical. ACF-Empirical represented each individual through sample autocorrelations at lags 1 through 10,

$$x_i = (\hat{\rho}_i(1), \hat{\rho}_i(2), \dots, \hat{\rho}_i(10)).$$

It summarized observed serial dependence without fitting an autoregressive order. Its coordinates differ in estimation precision and are not independent. In later analyses it is distinguished from model-implied and residual autocorrelations.

Feature-PCA. Feature-PCA applied principal components analysis to a prespecified set of 14 engineered features: the mean, standard deviation, skewness, excess kurtosis, and 25th, 50th, and 75th percentiles; sample autocorrelations at lags 1 through 5; the slope of a linear time trend; and the dominant frequency index from a fast Fourier transform. PCA reduced these features to two retained components.

Because series were z-normalized before feature extraction, the mean was zero and the standard deviation was one to within numerical precision. Thus, two features carried no discriminating information by construction. Features were standardized before PCA, and retained component scores were standardized again before clustering. This was the thesis's

implementation rather than a general recommendation. PCA retains directions of feature variance without using subgroup labels (Jolliffe & Cadima, 2016), so those directions need not support recovery. This construction represents one implementation of the feature-based family (Fulcher & Jones, 2017).

CNN-Pred. CNN-Pred used the autoregressive coefficients predicted by a convolutional network. It had one coordinate in AR(1) conditions and two in AR(2) conditions,

$$x_i = (\hat{\varphi}_i^{\mathrm{CNN}})$$

or

$$x_i = (\hat{\varphi}_{1i}^{\mathrm{CNN}}, \hat{\varphi}_{2i}^{\mathrm{CNN}}).$$

Its dimension matched Model-Based, but its estimates came from a learned mapping.

CNN-Latent. CNN-Latent used the 32-dimensional penultimate-layer activations of the same network, which support coefficient prediction without a parameter interpretation for each coordinate. CNN-Pred and CNN-Latent are therefore two representations from one coefficient-targeted architecture. They differ in both information retained and feature-space geometry, because the output layer compresses the latent representation.

CNN Architecture and Training. The CNN used two one-dimensional convolutional blocks. The first applied 32 filters with kernel width 7 and padding 3. The second applied 64 filters with kernel width 5 and padding 2. Each block included rectified linear unit activation, max pooling by a factor of two, and dropout with probability .50.

The resulting activations were flattened and passed to a fully connected layer with 32 units, followed by a rectified linear unit activation and dropout with probability .50. These activations formed the latent representation. A final output layer predicted one coefficient for AR(1) or two for AR(2), using a hyperbolic tangent activation. This architecture followed the

general use of convolutional feature extraction for sequential inputs described in the neural-network and time-series learning literature (LeCun et al., 1998; Ismail Fawaz et al., 2019).

The hyperbolic tangent activation constrained each predicted coefficient to (−1, 1). This bound was an architectural choice rather than a complete stationarity constraint. In particular, stationary AR(2) processes can have a first coefficient outside that interval, and independently bounded coefficient predictions need not jointly satisfy the AR(2) stationarity conditions. The restriction could therefore affect both prediction and the distribution supplied to the clustering algorithm.

Training followed the settings described in Section 5. Five independently seeded networks were trained per condition.

Training series were generated separately from the Monte Carlo evaluation datasets using the same condition-specific process and subgroup-centroid configuration. CNN targets were estimated by OLS from the observed training series using a vectorized least-squares routine rather than the statsmodels fit used for Model-Based. Targets were bounded to [−0.999, 0.999] in AR(1) conditions and [−1.5, 1.5] in AR(2) conditions. Model-Based and CNN-target estimates therefore differed in implementation and bounds. True generating coefficients were retained for diagnostics but were not training targets. Membership labels were not supplied as classification targets, although the training distribution reflected the generating population.

Each Study 1 network was trained on the number of series specified by its condition, so training-sample size varied with *N*. Study 2 networks instead used 500 training series each. Training-sample size therefore differed between studies, but supervision type did not.

This was therefore a supervised coefficient-estimation benchmark using estimated rather than generating coefficients as training targets. The network learned a mapping from observed

time series to OLS-estimated dynamic parameters using independent simulated training data. Its performance remained specific to the evaluated architecture, optimization procedure, training distribution, and coefficient estimator.

Once trained, network weights remained fixed and dropout was disabled during evaluation. For each replication, the observed series were passed through each trained network to obtain CNN-Pred and CNN-Latent. Each representation was clustered independently under the GMM and *k*-means procedures, scored against the generating memberships, and included in the within-replication seed average described in the Analysis Plan.

6.2 Results

Overall Method Comparison. Study 1 produced a clear overall ranking across the 144 clustering conditions. Model-Based achieved the highest pooled mean ARI (M = .290, MCSE = .001; *k*-means), followed by CNN-Pred (M = .212, MCSE < .001; GMM), CNN-Latent (M = .204, MCSE < .001; GMM), ACF-Empirical (M = .165; *k*-means), and Feature-PCA (M = .153; GMM). Under the primary within-condition best-algorithm definition, Model-Based had the highest ARI in 114 of 144 conditions (79.2%). This share was 76.4% under the diagonal GMM alone and 93.1% under *k*-means alone, showing that the magnitude of its advantage depended partly on the downstream procedure. Table 1 summarizes pooled performance and the condition-level counts.

A clean diagnostic condition examined whether the implemented pipelines could recover clearly separated structure under favorable conditions. That condition used AR(1), N = 300, T = 100, and Easy separation, with no within-subgroup coefficient spread and no observation noise. Model-Based attained an ARI of .94, followed by CNN-Pred (.90), CNN-Latent (.70), Feature-PCA (.45), and ACF-Empirical (.40). These results demonstrate that some pipelines achieved

high recovery under favorable conditions, although the diagnostic does not establish why performance was lower across the broader factorial design.

Although CNN-Pred learned to predict OLS estimates from independent, condition-matched training series, Model-Based achieved higher pooled recovery through direct coefficient estimation. This finding remains specific to the evaluated neural architecture, optimization procedure, and training distribution.

Condition-Level Performance. Condition-level counts allowed each representation its better clustering algorithm within each condition. Model-Based had the highest ARI in all 72 conditions with $N = 100$. The 30 conditions in which it did not have the highest ARI occurred entirely at $N = 300$ and became less frequent as series length increased: 13 at $T = 15$, 11 at $T = 50$, four at $T = 100$, and two at $T = 200$. CNN-Latent had the highest ARI in 22 of these conditions, and CNN-Pred and ACF-Empirical in four each. This pattern is consistent with larger samples providing more training information for the deep learning representations while direct coefficient estimation depends primarily on the information available within each person's series. Because bias and variance were not separately decomposed, this interpretation is treated as a plausible mechanism rather than a demonstrated explanation.

Condition-level inferential comparisons supported the overall ranking. The Friedman test rejected equal performance across the five representations, $\chi^2(4) = 289.57$, $p < .001$, Kendall's $W = .50$. Model-Based had the best mean rank (1.22), and Holm-corrected Wilcoxon signed-rank comparisons showed that it exceeded each competitor across the evaluated conditions (all $p < .001$). Median per-condition advantages were .068 over ACF-Empirical, .056 over CNN-Pred, .062 over CNN-Latent, and .090 over Feature-PCA, with matched-pairs rank-biserial correlations ranging from .88 to 1.00.

Effects of the Experimental Factors. The descriptive factor decomposition showed that Model-Based recovery was governed primarily by within-person information and subgroup distinctiveness. Under the diagonal GMM, series length accounted for 50.7% of the variation in condition-level mean ARI, subgroup separation for 31.4%, and signal-to-noise ratio for 7.4%. Sample size and temporal order contributed comparatively little, accounting for 0.1% and 0.3%, respectively. The largest interaction was $T \times$ Separation, which accounted for 6.7% of the variation. The remaining reported two-way interactions were approximately 1% or less. The full hierarchy is reported in Table 2.

Recovery improved strongly with series length. Under *k*-means, its pooled-best clustering algorithm, mean Model-Based ARI increased from .065 at $T = 15$ to .236 at $T = 50$, .369 at $T = 100$, and .490 at $T = 200$. At $T = 15$, all representations were near the floor, indicating that representation choice had limited value when individual series contained very little temporal information.

Subgroup separation and observation noise produced the expected ordering. Under the same algorithm, mean Model-Based ARI was .441 under Easy separation, .279 under Medium, and .150 under Hard. It increased from .213 at SNR = 1 to .298 at SNR = 2 and .359 at SNR = 4. The $T \times$ Separation interaction reflects that the benefit of longer series was largest when subgroups were well separated and smallest when they were not.

Representation Form and Recovery. CNN-Pred and CNN-Latent were derived from the same coefficient-targeted network but presented its learned information to the clustering algorithm differently. CNN-Pred operated in the same one- or two-dimensional coefficient space as Model-Based, whereas CNN-Latent used a 32-dimensional internal representation. Their

pooled means were similar (.212 and .204), but CNN-Latent accounted for most of the conditions in which Model-Based did not have the highest ARI.

This comparison illustrates a distinction that becomes more important in the Unknown-K analysis and the empirical demonstration: the information retained by a representation and the coordinate structure through which that information is presented to the clustering algorithm are separate properties.

Unknown-K Behavior. The primary Study 1 analysis fixed $K = 3$ to evaluate representation quality independently of cluster-number estimation. The separate Unknown-K analysis allowed the number of components to vary from 1 to 4. It retained Study 1's factor structure and included 192 conditions: 48 no-subgroup conditions and 144 three-subgroup conditions. The analysis differed from Study 1 in four respects: 500 rather than 1,000 replications per condition; convolutional networks trained against generating coefficients rather than OLS estimates; one trained network per condition rather than five; and full-covariance rather than diagonal-covariance Gaussian mixtures. Consequently, the Unknown-K findings characterize a separate analysis rather than isolating the effect of allowing K to vary.

Cluster number was selected using the Bayesian information criterion (BIC) with the Gaussian mixture model for Model-Based, ACF-Empirical, ACF-Implied (the ten model-implied autocorrelations computed from the Model-Based coefficient estimate), Feature-PCA, and CNN-Pred. For CNN-Latent, it was selected using the gap statistic (Tibshirani et al., 2001). Each selection procedure therefore pairs a criterion with a clustering algorithm.

False-positive rates under the 48 no-subgroup conditions differed sharply across representations. Model-Based had the lowest rate at 6.0%, followed by CNN-Latent at 8.1%, ACF-Empirical at 18.4%, CNN-Pred at 54.4%, Feature-PCA at 64.6%, and ACF-Implied at

100.0%. The contrast between Model-Based and ACF-Implied is especially informative because both were constructed from the same fitted autoregressive information. ACF-Implied transformed that information into ten model-implied autocorrelations and selected more than one component in every null replication. Thus, adding deterministic coordinates without adding fitted dynamic information substantially changed cluster-number calibration under the evaluated full-covariance mixture and BIC.

The gap statistic reduced, but did not remove, ACF-Implied's inflation, yielding a false-positive rate of 67.7%. CNN-Pred showed a different calibration problem: its BIC false-positive rate increased from 12.5% at $N = 100$ to 96.2% at $N = 300$. Using the gap statistic reduced this false-positive rate to 14.7%, but its correct $K = 3$ selection rate fell to 4.1%, illustrating a trade-off between false-positive control and sensitivity.

Correct $K = 3$ selection rates varied across representations under their primary selection procedures: 3.9% for Model-Based, 2.7% for ACF-Empirical, 21.7% for ACF-Implied, 12.2% for Feature-PCA, 8.9% for CNN-Latent, and 38.8% for CNN-Pred. No evaluated representation combined consistently low null false-positive rates with high correct-count selection. Table 3 reports the complete comparison.

These findings reinforce the distinction between recovering memberships when K is supplied and determining how many components should be selected. They also support retaining fixed $K = 3$ for the primary recovery comparisons, where the objective was to isolate representation quality rather than combine it with representation-specific cluster-number calibration.

7 Study 2: Mixed Temporal Order

7.1 Method

Changes From Study 1. Study 2 extended the comparison to populations containing individuals with different temporal orders. Whereas every individual within a Study 1 dataset followed the same autoregressive family, Study 2 included both AR(1) and AR(2) individuals within every generating subgroup. The objective was to evaluate subgroup recovery when the intended subgroups shared a persistence-based distinction that crossed temporal order.

Study 2 also introduced changes in representation construction that affect how its results can be compared with Study 1. As in Study 1, the CNN was trained on OLS-estimated coefficients from independently generated training series. However, Study 2 used targets from the common AR(2) fitting specification applied across its mixed-order population. Feature-PCA also used post-fit characteristics rather than the descriptive features used in Study 1 and retained ten principal components rather than two.

The studies additionally differed in population structure, subgroup definition, autoregressive fitting specification, CNN training-sample size, the evaluated range of series lengths, and replication count. Consequently, cross-study differences cannot be attributed to temporal-order heterogeneity alone.

The general evaluation procedures remained unchanged. Each representation was applied to the same evaluation datasets, both the diagonal-covariance GMM and *k*-means were supplied with $K = 3$, and ARI measured agreement with generating memberships. Five independently trained networks were evaluated per condition, with their ARI scores averaged within each replication.

Persistence-Space Construction. A mixed-order design must distinguish recovery of the intended subgroups from recognition of temporal order. If AR(1) individuals formed one subgroup and AR(2) individuals another, a representation could recover memberships by identifying order alone. Study 2 avoided that construction by including both orders within each subgroup and defining subgroup differences through shared persistence targets.

For a stationary AR(1) process, the lag-1 autocorrelation equals its autoregressive coefficient,

$$\rho_1 = \varphi_1.$$

For a stationary AR(2) process, the corresponding relationship is

$$\rho_1 = \varphi_1 / (1 - \varphi_2).$$

This relationship allows processes with different temporal orders to have a common target for one-step persistence while differing in their second-lag structure.

Each subgroup *k* was assigned a persistence target ρ_{1k} and a second-coefficient target φ_{2k} for its AR(2) members. For an AR(1) individual i assigned to subgroup *k*, the generating coefficient was

$$\varphi_{1i} = \rho_{1k} + u_{1i}.$$

For an AR(2) individual in the same subgroup, coefficients were generated as

$$\varphi_{2i} = \varphi_{2k} + u_{2i},$$

and

$$\varphi_{1i} = \rho_{1k}(1 - \varphi_{2i}) + u_{1i},$$

where the independent perturbations followed

$$u_{1i}, u_{2i} \sim N(0, \tau^2), \quad \tau = 0.10.$$

The construction aligned subgroup persistence centers across orders without requiring identical within-subgroup persistence distributions. Substituting the AR(2) coefficient construction into the autocorrelation identity gives

$$\rho_{1i} = \rho_{1k} + u_{1i} / (1 - \varphi_{2i}).$$

Thus, the realized persistence spread among AR(2) members depended on their second coefficients. The perturbation standard deviation τ did not imply a common persistence standard deviation across orders. Stationarity restrictions further affected the accepted coefficient distributions. In generation, AR(1) coefficients were clipped to the interval $[-0.99, 0.99]$, whereas AR(2) coefficient pairs were obtained by rejection sampling under the stationarity constraints, with an error raised if no admissible pair was found within 1,000 attempts.

Subgroup Separation and Order Composition. The Easy, Medium, and Hard conditions used the following sets of (ρ_1, φ_2) target pairs, respectively: {(0.10, 0.00), (0.50, 0.20), (0.75, −0.10)}, {(0.20, 0.00), (0.50, 0.20), (0.70, −0.10)}, and {(0.30, 0.00), (0.50, 0.20), (0.65, −0.10)}.

The first coordinate defined the shared persistence target. The second specified additional dynamics for AR(2) members, while AR(1) members retained $\varphi_2 = 0$. Subgroup membership therefore represented a persistence-centered distinction across orders rather than membership in a single autoregressive family.

Three order compositions were evaluated: AR1-Dominant, containing 70% AR(1) individuals; Mixed, containing equal proportions of AR(1) and AR(2) individuals; and AR2-Dominant, containing 30% AR(1) individuals. The condition-specific composition was applied within every subgroup, subject to integer allocation. Generating subgroups were therefore not distinguished by their intended order proportions.

Experimental Design. Study 2 crossed order composition, subgroup separation, series length ($T \in \{50, 100, 200\}$), signal-to-noise ratio (SNR $\in \{1, 2, 4\}$), and sample size ($N \in \{100, 300\}$).

The resulting design contained 162 conditions, each replicated 500 times. Subgroup sizes were balanced as closely as integer allocation permitted.

The burn-in procedure, observation-noise construction, and general preprocessing followed Study 1. The shortest Study 1 series length, $T = 15$, was not included. Study 2 therefore evaluated the mixed-order comparison over the longer series lengths rather than extending the comparison to every Study 1 condition.

A common AR(2) model was fitted to every individual for representations requiring estimated coefficients. At the latent-process level, this specification accommodates AR(1) as the special case $\varphi_2 = 0$. It nevertheless requires estimation of an additional coefficient for true AR(1) individuals. The common specification therefore avoids switching coefficient dimensions across people while introducing an estimation cost when the second coefficient is unnecessary. As in Study 1, additive observation noise can affect the relationship between fitted and generating coefficients.

Representations Compared. Study 2 compared the same five representations, constructed as follows.

Model-Based. Every individual was represented by the OLS coefficient pair from the common AR(2) fit,

$$x_i = (\hat{\varphi}_{1i}, \hat{\varphi}_{2i}).$$

The representation had two coordinates regardless of generating order. True AR(1) individuals were not assigned a second coefficient of zero by construction. Instead, their second

coefficient was estimated from the observed series. The common AR(2) model was fitted by least squares on the lagged design matrix without an intercept term.

ACF-Empirical. Each individual was represented by the sample autocorrelations at lags 1 through 10. This was the same empirical-autocorrelation construction used in Study 1. It did not require selecting a temporal order for each person.

Feature-PCA. The engineered representation was constructed from ten post-fit features: the two estimated AR coefficients, the residual mean square, the Akaike information criterion (Akaike, 1974), the Bayesian information criterion (Schwarz, 1978), the Gaussian log-likelihood, and the mean, standard deviation, skewness, and excess kurtosis of the residuals. Nonfinite feature values were replaced with the corresponding feature median across individuals within the dataset. The features were standardized before principal components analysis, and the retained component scores were standardized again before clustering. Study 2 inherited the configured default of ten components and therefore retained ten component scores rather than reducing the ten engineered features to a smaller number.

This study-specific implementation combined fitted temporal information with characteristics of model fit and residual behavior, unlike Study 1's descriptive feature set. Study 1 retained two principal components, whereas Study 2 retained the configured default of ten. Because the feature sets and retained dimensions differed without an isolated dimensionality manipulation, the Feature-PCA results should not be compared across studies as though dimensionality alone had been varied experimentally.

CNN-Pred. The CNN used the architecture described in Study 1 but predicted the two OLS AR(2) coefficients. Its training targets were obtained by fitting the common AR(2) model to independently generated training series. Each network was trained on 500 series per condition,

so increasing the evaluation sample from $N = 100$ to $N = 300$ did not increase its training-sample size.

The network therefore learned a mapping to estimated coefficients rather than to the generating parameters. This provided a target available without knowledge of true coefficients, while incorporating the properties of OLS estimation into the supervision.

CNN-Latent. The 32-dimensional penultimate layer of the same trained network provided CNN-Latent. As in Study 1, CNN-Pred and CNN-Latent were evaluated separately. Network weights remained fixed during Monte Carlo evaluation, and scores were averaged across the five training seeds within each replication.

Order-Decoupling and Prediction Diagnostics. Two diagnostics assessed whether recovered partitions primarily distinguished temporal order rather than the intended subgroups. Order-ARI compared recovered memberships with the generating AR(1)/AR(2) labels. A second diagnostic summarized the largest deviation between a recovered cluster's AR(1) proportion and the population composition.

These diagnostics addressed alignment between the partition and order. Low order-ARI would support interpreting subgroup ARI against the intended grouping target, but it would not demonstrate that the representation contained no information about temporal order.

Held-out CNN diagnostics also compared predicted coefficients with their OLS targets. Correlations described association with the targets, while ratios of predicted to target standard deviations described the spread of the predictions. These quantities were used to characterize the learned mapping rather than to establish unbiasedness or recovery of true coefficients.

7.2 Results

Order-Decoupling Results. The recovered partitions showed little evidence of primarily separating individuals by temporal order. Pooled order-ARI values were within ±.003 of zero for every representation under both clustering algorithms, and the largest reported condition-level mean was .036. Deviations in recovered-cluster order composition ranged from .08 to .15. These diagnostics support evaluating recovery against the intended persistence-based memberships, although they do not imply that the representations contained no information about temporal order.

Overall Method Comparison. Model-Based and the two CNN-derived representations performed much more similarly in Study 2 than in Study 1. Under the diagonal GMM, pooled mean ARI was .2515 for CNN-Latent, .2458 for CNN-Pred, and .2391 for Model-Based, a range of only .0124. Under *k*-means, the ordering changed: CNN-Pred led at .2417, followed by Model-Based at .2384 and CNN-Latent at .2352, a range of .0065. ACF-Empirical and Feature-PCA trailed the leading three under both clustering algorithms. Monte Carlo standard errors of these pooled means were .0004 or smaller, so the .0124 range among the leading three exceeds Monte Carlo error, although the differences are small. Feature-PCA's low recovery may partly reflect that all ten standardized component scores were retained, so components carrying little subgroup information received the same weight as those carrying most of it; this explanation was not isolated. Table 4 reports the complete comparison.

Under the within-condition best-algorithm definition, CNN-Latent had the highest ARI in 75 of 162 conditions (46.3%), Model-Based in 47 (29.0%), and CNN-Pred in 40 (24.7%). Neither ACF-Empirical nor Feature-PCA had the highest ARI in any condition. Thus, in Study 2 the three leading representations performed similarly, and their ordering depended on the

clustering algorithm and the information available in each condition rather than on a single representation that dominated the design.

The Friedman test rejected equal performance, $\chi^2(4) = 509.07$, $p < .001$, Kendall's $W = .79$. Mean ranks placed CNN-Latent first (1.67), CNN-Pred second (2.05), and Model-Based third (2.39), ahead of ACF-Empirical (3.90) and Feature-PCA (5.00). Holm-corrected Wilcoxon comparisons showed that Model-Based trailed CNN-Latent by a median .016 per condition and CNN-Pred by .007, while exceeding ACF-Empirical by .060 and Feature-PCA by .167. The differences among the leading three were therefore statistically consistent across the design but modest in absolute magnitude and sensitive to the downstream clustering algorithm.

Effects of the Experimental Factors. For Model-Based under the diagonal GMM, series length and subgroup separation dominated the descriptive factor decomposition, accounting for 43.5% and 40.6% of variation in condition-level mean ARI, respectively. Signal-to-noise ratio accounted for another 12.1%, whereas order composition and sample size accounted for only 0.9% and 0.6%. The reported interactions were small: $T \times$ Separation accounted for 0.8%, Separation × SNR for 0.3%, and $T \times$ Composition for 0.1%. Table 5 reports the full decomposition.

These percentages should not be compared directly with Study 1 because the designs span different ranges of T. In particular, Study 2 excludes $T = 15$, where Study 1 recovery was near the floor. Within Study 2, series length, subgroup separation, and observation noise accounted for substantially more variation in Model-Based recovery than order composition or sample size.

Series length also changed the relative ranking of the leading representations. Model-Based had the highest ARI in 11 of 54 conditions at $T = 50$, 13 at $T = 100$, and 23 at $T = 200$,

corresponding to shares of 20.4%, 24.1%, and 42.6%. CNN-Latent showed the opposite pattern, having the highest ARI in 53.7%, 51.9%, and 33.3% of conditions at the three series lengths. CNN-Pred's share was comparatively stable across T. Even at $T = 200$, however, the two deep learning representations together had the highest ARI in 57.4% of conditions. Longer series therefore made direct coefficient estimation increasingly competitive without producing a complete reversal in favor of Model-Based. Figure 1 plots these shares across series lengths.

The size of this longer-series advantage depended on the clustering algorithm. At $T = 200$, Model-Based had the highest ARI in 30 of 54 conditions under k-means but 21 under the diagonal GMM. This information-regime pattern, in which the leading representation depends on how much information each series carries, should therefore be interpreted as a property of representation-and-algorithm combinations rather than of representation alone.

Coefficient Prediction and Representation Contraction. Held-out CNN diagnostics showed that predicted coefficients closely tracked the OLS targets while being substantially less dispersed. Correlations with the first- and second-coefficient targets were .944 and .834, whereas the standard deviations of the predictions were only 0.52 and 0.40 times those of the corresponding targets.

This contraction provides one plausible explanation for the deep learning representations' relative advantage in shorter-series conditions. Shrinking noisy person-level estimates toward the training distribution may reduce sampling variation presented to the clustering algorithm. However, contraction can also remove meaningful between-person variation, and these diagnostics compared predictions with estimated rather than true coefficients. The results therefore suggest, but do not establish, a bias-variance or regularization mechanism. Direct comparison with explicitly regularized estimators would be required to test that explanation.

8 Robustness Analyses, Extensions, and Diagnostics

The primary studies evaluated subgroup recovery when membership was defined by autoregressive dynamics. This section reports three kinds of additional analysis. Robustness analyses examined whether the representation ordering held when the number of generating subgroups or the latent width of the network changed. Extensions addressed further questions: how the model-fitting strategy applied before clustering affects recovery, and how the ranking changes when scale or conditional volatility rather than persistence defines subgroup membership. A diagnostic examined recovery once within-subgroup coefficient spread and additive observation noise were removed. Because several analyses changed the scientific target rather than merely perturbing the primary design, their results are interpreted as boundary conditions rather than as a single extended ranking.

8.1 Extension: Selection of Temporal Order

This extension used a separate mixed-order design to examine three fitting strategies: a common AR(1) fit, a common AR(2) fit, and person-specific BIC-based order selection. In this design, subgroups differed in the first autoregressive coefficient, with the same centroids as Study 1's AR(1) conditions ({0.1, 0.5, 0.9}, {0.2, 0.5, 0.8}, and {0.3, 0.5, 0.7} for Easy, Medium, and Hard separation) and a perturbation standard deviation of $\tau = 0.10$. Within each subgroup, 70%, 50%, or 30% of individuals followed an AR(1) process, and the rest followed an AR(2) process whose second coefficient was drawn around -0.30 ($\tau = 0.10$, subject to stationarity). Six representations were evaluated with a Gaussian mixture: Model-Based, ACF-Implied, ACF-Residual, Feature-PCA, CNN-Pred, and CNN-Latent. The producing script specifies a full-covariance mixture, which applies to at least 126 of the 162 conditions; the setting used for the remaining 36, which were run with an earlier version of the script, is not

recorded. It is an extension rather than a robustness analysis of the primary Study 2, because it used a different subgroup definition. The primary Study 2 remains the mixed-order persistence-space design reported above.

Under the common AR(1) strategy, Model-Based achieved the highest pooled mean ARI (.314) and had the highest ARI in 138 of 162 conditions (85.2%), followed by ACF-Implied (.243), CNN-Pred (.228), Feature-PCA (.223), CNN-Latent (.178), and ACF-Residual (.059). Under the common AR(2) strategy, Model-Based again had the highest pooled mean (.239), although CNN-Pred (.216) led more frequently than under AR1-fit. These results show that a simpler common representation can yield better subgroup recovery in this design even when it is misspecified for some individuals. One plausible explanation is that estimating additional person-specific parameters introduces variation that describes individual dynamics without improving separation of the generating subgroups. However, the comparison does not isolate this mechanism.

The factor decomposition for Model-Based under AR1-fit (Table 6) placed subgroup separation first ($\eta^2 = .589$), followed by series length (.227) and signal-to-noise ratio (.148). This ordering differs from the persistence-space decomposition in Table 5 because the two analyses use different subgroup definitions and representations rather than repeated estimates of one design.

Person-specific BIC-based order selection performed substantially worse. Under BIC-Adaptive, the highest pooled mean ARI among the six representations was only .078. For Model-Based, mean ARI fell from .314 under AR1-fit to .064 under BIC-Adaptive. The adaptive pipeline combined errors in order selection with a mixed-dimensional representation in which AR(1)-selected individuals were zero-padded in the second coefficient. Distances could therefore

reflect the routing decision as well as the subgroup-defining dynamics. This finding applies to the implemented BIC-Adaptive pipeline rather than to adaptive modeling in general, but it demonstrates that the fitting strategy used before clustering can fundamentally alter the recovery problem.

The contrast between ACF-Implied and ACF-Residual further showed that representation usefulness depended on where subgroup-relevant information remained after fitting. Under AR1-fit, ACF-Implied exceeded ACF-Residual by .184 ARI on average, indicating that the fitted dynamics carried considerably more of the generating subgroup information than residual dependence in this design.

8.2 Robustness to Subgroup Count

A matched subset of Study 2 conditions was rerun with $K = 2$ and $K = 4$ and compared with $K = 3$ while supplying the true count to the clustering algorithm. Under the diagonal GMM, recovery of the four representations other than Feature-PCA decreased as the number of subgroups increased (Feature-PCA stayed low, .018 to .051), but the ordering of the five primary representations remained unchanged across K. The information-regime pattern also persisted: at $K = 2$, CNN-Latent exceeded Model-Based at $T = 50$, whereas Model-Based exceeded CNN-Latent at $T = 200$. These results indicate that the primary ranking pattern was not unique to three generating subgroups, while remaining specific to the matched subset examined and to the diagonal Gaussian mixture.

8.3 Subgroup-Defining Signal as a Boundary Condition

Two analyses changed the property defining subgroup membership. In the scale-shift analysis, subgroups differed in series scale while persistence was held constant and within-person standardization was disabled. A scale-retaining Feature-PCA representation achieved the

highest pooled mean ARI (.560) and had the highest ARI in 67% of conditions, whereas Model-Based and ACF-Empirical were approximately zero. Thus, the advantage of direct autoregressive representations did not extend to subgroups defined by information those representations do not retain.

An eight-condition volatility pilot produced the same conceptual result for conditional variance. A volatility-specific generalized autoregressive conditional heteroskedasticity (GARCH) representation attained the highest recovery in all eight conditions, with pooled mean ARI .185, whereas Model-Based and ACF-Empirical were near zero. Absolute recovery remained modest, so the pilot should not be interpreted as a general comparison of volatility methods. Together, the scale and volatility analyses show that representation performance depends fundamentally on alignment between the representation and the property defining subgroup membership.

8.4 Reduced-Noise Diagnostic and Latent-Width Robustness Check

A reduced diagnostic crossed 24 base conditions with four data-generating variants, yielding 96 condition–variant combinations. The variants included a baseline retaining both sources of variation and three alternatives that removed within-subgroup coefficient variation, additive observation noise, or both. Under the common AR(1) fitting strategy, Model-Based was the strongest representation under every variant, and its mean ARI increased from .401 in the reduced design's baseline to .685 when both sources were removed. That comparison is internal to the reduced diagnostic and should not be read as a change from the fitting-strategy extension's pooled mean of .314, which comes from a different design. The remaining gap from perfect recovery may reflect finite-series sampling variation, stochastic innovations, model fitting, and clustering error. The diagnostic does not identify their separate contributions.

Implementation checks also indicated that the CNN findings were not driven by one uniquely favorable latent width. CNN-Pred performance varied more at smaller widths but was comparatively similar across widths from 32 to 40, while CNN-Latent achieved its highest reported mean at width 32. These checks support robustness to the evaluated architectural variation without establishing optimization adequacy or general robustness across neural architectures.

9 Empirical Demonstration: How Nuts Are the Dutch

The empirical demonstration applied the representations from the simulations to experience-sampling affect data from the How Nuts Are the Dutch project. It adapted the procedure that Li et al. (2025) used with data from this project, in which person-specific vector autoregressive (VAR) models of positive and negative affect are estimated, their coefficients are clustered, the number of clusters is selected, and the resulting clusters are described by their dynamics and by external measures. The five representations from the simulations were added to Li et al.'s two VAR representations and analyzed with the same steps.

Because the data provide neither known memberships nor known generating parameters, the demonstration does not evaluate recovery accuracy, and agreement between solutions is not interpreted as accuracy. Its purpose is to show how the selected number of clusters, the resulting memberships, and the cluster-level dynamics depend on the representation and the selection procedure when the pipelines are applied to real data. It does not aim to establish subgroups in this population.

9.1 Method

Data and Sample. The data came from the diary study of the How Nuts Are the Dutch project, a Dutch crowdsourcing study of mental symptoms and strengths in which participants

from the general population reported their momentary affect three times a day (morning, midday, and evening) for 30 days, giving 90 scheduled occasions (van der Krieke et al., 2016, 2017). Data from this project have been used in previously published studies, including the VAR-based clustering application of Li et al. (2025). The file available for this thesis contained 115,386 rows from 1,396 participants (Li et al. report 1,393) and 1,550 diary episodes. Each episode was restricted to its first 90 scheduled occasions, and episodes without the complete 90-row record were excluded (1,165 episodes remained). For each participant, the episode with the most occasions at which all 12 affect items were answered was retained (1,072 participants), and at least 81 such occasions were required. The final sample comprised 178 participants with a mean of 83.9 completed occasions (range 81 to 90). The sample therefore represents participants with nearly complete diaries rather than the full study population.

Measures. At each occasion, participants rated 12 affect adjectives from 0 to 100: six positive (enthusiastic, energetic, cheerful, relaxed, content, and calm) and six negative (irritable, nervous, anxious, tired, dull, and gloomy; Li et al., 2025). Following the original analysis as coded, positive affect (PA) and negative affect (NA) were each represented by a single diary item: item 5 for PA and item 8 for NA. Both items were answered at 14,935 occasions (M = 83.9 per participant). Before the diary, participants completed measures of depression (the Quick Inventory of Depressive Symptomatology [QIDS] and the depression scale of the Depression Anxiety Stress Scales [DASS-21]), anxiety and panic, and positive and negative affect. These pre-diary measures were used only to describe the resulting clusters, as Li et al. described their clusters with external measures. The DASS-21 was available for 103 participants and the other measures for all 178.

Time Grid and Imputation. Occasions were scheduled roughly six hours apart during the day and twelve hours apart overnight. As in Li et al. (2025), each participant's reports were placed on a six-hour grid with an empty position after each evening report, so that the overnight interval counted as two steps. This gave 120 grid positions per participant, 90 of them scheduled occasions. About 30% of the 21,360 grid positions were missing, including all 5,340 overnight positions. Missing values were multiply imputed (m = 10) with a single imputation model fitted to all participants together. The model used PA, NA, their preceding and following values, and the position within the day, and it drew imputed values from its predictive distribution (the iterative imputer of scikit-learn; Pedregosa et al., 2011). Li et al. used a different imputation program. Imputed values were not restricted to the response scale. Each participant's series was centered on that participant's mean before models were fitted.

Person-Level Dynamics. For each participant and imputed data set, VAR(1) and VAR(3) models of PA and NA without intercepts were estimated by least squares, and the coefficients were averaged over the 10 imputations. By the Akaike information criterion (Akaike, 1974), VAR(1) was the best-fitting order for 121 participants, VAR(2) for 31, and VAR(3) for 26 (Li et al., 2025, reported 144, 46, and 37 of 227). As in Li et al., VAR(1) and VAR(3) were then fitted to every participant as the lower- and higher-order representations, with 4 and 12 coefficients per participant. The mean person-specific VAR(1) coefficients were 0.245 (SD = 0.099) for PA on preceding PA, −0.023 (SD = 0.118) for PA on preceding NA, −0.012 (SD = 0.057) for NA on preceding PA, and 0.238 (SD = 0.165) for NA on preceding NA.

Representations. The five representations from the simulations were constructed from the same imputed series and averaged over imputations in the same way. Model-Based was the AR(1) coefficient of each affect, two values per participant; unlike VAR(1), it omits the cross-

lagged coefficients through which each affect predicts the other. ACF-Empirical was the sample autocorrelations at lags 1 to 10 of each affect (20 values). Feature-PCA used 13 engineered features of each affect (standard deviation, skewness, kurtosis, three quartiles, autocorrelations at lags 1 to 5, linear slope, and dominant frequency), standardized and reduced to two principal components that accounted for 43.8% of the variance. For CNN-Pred and CNN-Latent, the convolutional architecture used in the simulations, with PA and NA as two input channels, was trained to predict each participant's four VAR(1) coefficients from that participant's within-person standardized series (1,780 series, 178 participants × 10 imputations; 150 epochs; Adam optimizer, learning rate .001; mean squared error loss). CNN-Pred was the four predicted coefficients and CNN-Latent the 32-unit internal embedding, of which five units were zero for every participant, leaving 27 values. The predicted coefficients correlated .89 to .96 with the estimated VAR(1) coefficients. Unlike in the simulations, the network was trained on the same participants it then represented. As in Li et al. (2025), the VAR coefficients were clustered as estimated; the five representations were standardized before clustering.

Number of Clusters, Membership, and Cluster-Level Dynamics. Following Li et al. (2025), the number of clusters was selected from 1 to 8 with two selection procedures, each pairing a criterion with a clustering algorithm. The first used the Bayesian information criterion (BIC; Schwarz, 1978) with the Gaussian mixture model (GMM). BIC chose both the number of clusters and the covariance structure among spherical, diagonal, tied, and full structures, and each mixture was fitted from 20 random starts and one start from hierarchical clustering. The second used the gap statistic (Tibshirani et al., 2001) with average-linkage hierarchical clustering and 500 uniform reference data sets, selecting the number of clusters with the largest gap; *k*-means, started from the hierarchical clusters, then assigned memberships. Agreement between

the solutions of different representations was summarized by the adjusted Rand index (ARI; Hubert & Arabie, 1985). As in Li et al., memberships were taken from the VAR(1) solution under BIC with the GMM. The clusters were described by their size, age, and percentage of women; momentary PA and NA; the correlation between participant-mean PA and NA; the cluster means of the participants' VAR(3) coefficients; and the pre-diary measures. Differences in the pre-diary measures were tested with one-way analyses of variance and a Holm (1979) correction across the five measures.

9.2 Results

Number of Clusters. Table 7 reports the number of clusters selected for each representation. With BIC and the GMM, two clusters were selected for VAR(1) (93 and 85 participants), VAR(3) (89 and 89), and CNN-Pred (98 and 80). In each of these solutions BIC preferred diagonal covariance matrices, and the selected number led the next-best number by 19 to 41 BIC points. One cluster was selected for Model-Based, ACF-Empirical, and Feature-PCA, although for Feature-PCA one and two clusters differed by only 0.2 BIC points. Model-Based omits the cross-lagged coefficients that VAR(1) and CNN-Pred include, which is one plausible reason it selected one cluster where they selected two. For CNN-Latent, BIC selected seven clusters, but this count is not interpretable: because many latent values were exactly zero, BIC continued to improve as tight components were added. With the gap statistic and *k*-means, one cluster was selected for every representation. The selected number of clusters therefore depended on both the representation and the selection procedure.

Agreement Across Representations. The three two-cluster solutions largely grouped the same participants. CNN-Pred agreed with VAR(1) at ARI = .77 and with VAR(3) at .65, and the two VAR solutions agreed at .56. CNN-Pred's closest match was therefore VAR(1), the

representation whose coefficients the network was trained to predict, so part of this agreement is expected by construction. The cluster profiles were also similar. In both the VAR(1) and the CNN-Pred solution, one cluster reported more momentary NA (22.2 and 21.6 vs. 6.0 and 5.8 on the 0–100 scale), stronger carry-over of NA in the cluster-mean VAR(3) coefficients (0.36 and 0.35 vs. 0.10 and 0.10), and higher pre-diary depression on the QIDS (8.20 and 7.90 vs. 4.76 and 4.92).

Cluster Characteristics. Table 8 describes the two VAR(1) clusters. They were similar in size (93 and 85 participants) and in the percentage of women (82.8% and 81.2%), and Cluster 1 was somewhat younger (M = 40.1 vs. 45.1 years). During the diary, Cluster 1 reported more NA (M = 22.2 vs. 6.0) and somewhat less PA (M = 56.9 vs. 63.3). Within Cluster 1, participants with higher mean PA had clearly lower mean NA ($r = -.52$, $p < .001$), whereas this association was weak in Cluster 2 ($r = -.19$, $p = .077$).

Cluster-Level Dynamics. Table 9 reports the cluster means of the participants' VAR(3) coefficients; one step corresponds to six hours on the grid. In Cluster 1, NA carried over strongly from one occasion to the next (0.36), and higher NA predicted lower PA at the next occasion (−0.10). In Cluster 2, NA carried over weakly (0.10), and its lagged association with PA was small and positive (0.06). Carry-over of PA was similar in the two clusters (0.23 and 0.26), and all coefficients at lags 2 and 3 were small ($|b| \leq 0.05$).

Pre-Diary Measures. Table 10 compares the clusters on the pre-diary measures. Cluster 1 reported more depression on the QIDS (M = 8.20 vs. 4.76) and the DASS-21 (M = 12.58 vs. 9.67), more anxiety and panic (M = 1.19 vs. 0.51), more negative affect (M = 23.70 vs. 17.12), and less positive affect (M = 30.58 vs. 34.09). All five differences remained significant after

Holm correction (all $p < .001$). As in Li et al. (2025), the cluster with the more persistent NA also reported more depression and anxiety.

Sensitivity Checks. Two further runs bear on interpretation. With diagonal covariance structures only, the three two-cluster solutions were unchanged, whereas ACF-Empirical then selected five clusters and Feature-PCA two. In a run that represented PA and NA by the mean of their six items instead of single items (one diary per participant under a slightly broader eligibility rule, $N = 184$), BIC with the GMM selected one cluster for VAR(1), VAR(3), Model-Based, ACF-Empirical, and Feature-PCA, and the gap statistic selected one for every representation; CNN-Pred's four-cluster solution included a cluster of one participant, and CNN-Latent's BIC again kept improving. With single items under the same broader rule, two clusters were again selected for VAR(1) and VAR(3). The two-cluster solutions therefore depended on representing affect by single items.

Summary. The demonstration showed in real data a practical point of the thesis: the number of clusters found depended on the representation and on the selection procedure. BIC with the GMM selected two clusters for the two VAR representations and CNN-Pred but one for Model-Based, ACF-Empirical, and Feature-PCA, and the gap statistic with *k*-means selected one for every representation. Li et al. (2025) found the same contrast between procedures: more than one cluster with BIC and the GMM, and one with the gap statistic and *k*-means. The results therefore show that the count depends on the representation and the selection procedure, not that the population contains two subgroups.

Several features limit interpretation. No null controls were run for the demonstration; in the separate Unknown-K analysis, with its own settings, CNN-Pred with BIC selected more than one component in 54.4% of no-subgroup data sets (Table 3). The network was trained on the

same participants it represented, so its agreement with VAR(1) is partly by construction. In Cluster 2, momentary NA was near the floor of the scale, so its weak NA carry-over may partly reflect having little NA to carry over, and the pre-diary differences may partly reflect the clusters' difference in NA level rather than in dynamics. PA and NA were each measured by a single item, and the two-cluster solutions did not appear with six-item averages. Finally, the sample comprised participants with nearly complete diaries, and about 30% of grid positions, including every overnight position, were imputed.

10 Cross-Study Synthesis

Across the two primary studies, the robustness analyses and extensions, and the empirical demonstration, five conclusions emerge.

First, direct parameter estimation provided the strongest recovery when subgroup membership was defined by autoregressive dynamics and individuals shared a common temporal order. Model-Based led Study 1 despite comparison with CNN representations trained to predict OLS-estimated autoregressive coefficients from independent, condition-matched training series. This favors direct use of a compact coefficient representation in the evaluated order-homogeneous setting while remaining specific to the estimator, architecture, training procedure, and clustering methods examined here.

Second, mixed-order populations produced a more conditional ranking. In Study 2, Model-Based, CNN-Pred, and CNN-Latent performed similarly, with the deep learning representations performing better more often when individual series were shorter and Model-Based becoming increasingly competitive as series length increased. Because Study 1 and Study 2 also differed in the other respects set out in the Study 2 Method, this change cannot be attributed to order heterogeneity alone.

Third, the combined pattern is consistent with an information-regime account. When person-level temporal information is limited, deep learning representations may benefit from contracting noisy individual estimates toward patterns supported by the training distribution. As within-person information increases, direct estimation becomes more competitive because its sampling variability decreases. The present studies support this explanation as a plausible mechanism rather than a demonstrated bias-variance decomposition.

Fourth, the form in which information is presented to the clustering algorithm can matter even when the underlying fitted information is unchanged. Model-Based and ACF-Implied were deterministic representations of the same fitted autoregressive information but exhibited sharply different cluster-number behavior in the Unknown-K analysis under a full-covariance mixture, and the analysis did not establish the mechanism producing it. In the empirical demonstration, the selected number of clusters likewise differed across representations of the same affect series and across selection procedures. Representation geometry affects not only membership recovery but also calibration of subgroup enumeration.

Fifth, no representation family was uniformly superior because performance depended on what property defined subgroup membership. Autoregressive representations performed well when subgroups differed in persistence-related dynamics but failed when the subgroup-defining information was scale or conditional volatility that those representations did not preserve. The broader conclusion is therefore one of representation alignment: the useful representation is the one that captures the hypothesized subgroup-defining process in a form that can be estimated reliably and used appropriately by the downstream clustering procedure.

11 Discussion

The Cross-Study Synthesis identified the main empirical pattern across the two simulations: subgroup recovery was strongest when the feature representation preserved the relevant autoregressive signal in a compact and common-dimensional form. The Discussion interprets that pattern in terms of representation geometry, finite-sample information, fitting strategy, and implications for intensive longitudinal research. The central point is not that one method should be preferred in all time-series clustering problems, but that method performance depends on whether the constructed feature space aligns with the dynamic structure that defines the subgroups.

In order-homogeneous populations, and in the high-information regime of mixed-order populations, direct parameter-space representations provided the most stable basis for subgroup recovery. Alternative feature spaces and the network's internal embedding often retained useful dynamic information, and in the low-information regime of Study 2 the embedding led; but across the design, that information was less consistently organized in a way that supported mixture recovery, and (as the Unknown-K analysis shows) some ways of organizing the same information can induce additional mixture components that need not correspond to distinct generating populations. The fitting-strategy extension adds that the fitting strategy shared by all representations can alter the recovery problem itself. These findings suggest that time-series clustering should be evaluated as a full pipeline, including model fitting, representation construction, rank structure, scaling, and the clustering algorithm, rather than as a comparison of clustering algorithms alone.

11.1 Recovery Levels and Boundary Conditions

A useful way to interpret the simulation results is to distinguish conditions in which subgroup recovery was at least moderate from conditions in which the clustering problem was essentially unrecoverable. For this purpose, the bands Steinley (2004) describes for the adjusted Rand index serve as reference points: .90 or above indicates excellent recovery, .80 or above good, .65 or above moderate, and below .65 poor recovery (see also Li et al., 2025). The bands are descriptive conventions rather than tests. By them, the pooled Study 1 mean for Model-Based (.290) indicates poor recovery, whereas the clean baseline (.94) indicates excellent recovery, which underlines that recovery across the broader design was modest.

In Study 1, Model-Based exceeded .65 in 12 condition cells, all under Easy separation at $T = 200$ (SNR 2 and 4) or $T = 100$ (SNR 4 only), across both AR orders and sample sizes, with a maximum of .825. CNN-Pred did so in three cells and CNN-Latent in one, all at $T = 200$, Easy, SNR = 4, and the two feature-based representations never did. In Study 2, no condition reached .65 for any representation. The highest condition means were .521 for CNN-Latent and .511 for Model-Based, both at $T = 200$ under Easy separation and SNR = 4. Mixed-order populations with subgroups defined in a shared persistence space did not reach even moderate recovery anywhere in the design examined.

The complementary pattern is equally important. Several design regions produced poor subgroup recovery across all methods, indicating that method choice could not compensate for insufficient recoverable signal under those conditions. In Study 1, the clearest boundary conditions were Hard separation and $T = 15$. AR(2) conditions with Hard separation produced near-chance ARI for all methods across the evaluated series lengths and sample sizes. At $T = 15$, condition means ranged from .002 to .165 across representations and clustering algorithms,

indicating limited recovery across methods and reducing the substantive importance of their relative differences. These findings identify Hard separation as a practical limitation for the clustering pipelines and conditions examined, rather than establishing a universal boundary beyond which clustering is not viable. Similarly, Hard separation combined with low SNR produced near-floor recovery for every representation in Study 2. In the fitting-strategy extension, the highest pooled mean ARI across representations under the BIC-Adaptive strategy was .078.

Together, the high-performance and poor-performance conditions sharpen the substantive interpretation of the simulation. The main question is not simply which method had the highest average ARI, but when subgroup recovery became feasible at all. High recovery required both informative data and an appropriate representation of the dynamic signal. Conversely, when temporal information, subgroup separation, or representation coherence was insufficient, all methods deteriorated, and differences among methods became less substantively meaningful.

These recovery levels are lower than those Li et al. (2025) reported for VAR-based clustering, where the average ARI was .80 under the lower-order and .55 under the higher-order specification. Their simulations differed in ways that favor recovery: individuals within a cluster shared identical coefficients, so within-cluster variation came only from the innovations; no measurement noise was added; and series ranged from 50 to 1,000 occasions. The closest condition here is the clean baseline, which removed coefficient spread and observation noise and gave .94 for Model-Based. Li et al. also found the Gaussian mixture better than *k*-means under the lower-order specification (.88 vs. .73), whereas in Study 1 Model-Based was slightly higher under *k*-means (.290 vs. .280); their BIC also chose the covariance structure, whereas the mixtures here used a fixed diagonal structure.

11.2 Representation Structure and Subgroup Recovery

A central implication of the present findings is that successful time-series clustering depends as much on the geometry of the feature representation as on the clustering algorithm itself. All representations were evaluated under the same two clustering algorithms, so differences in recovery within either clustering algorithm reflect how each pipeline represented the individual time series before clustering, and differences between clustering algorithms show how much the representation's advantage depends on the downstream procedure. The strongest representation in order-homogeneous populations was the simplest one: estimated autoregressive coefficients. By contrast, empirical ACF vectors, engineered Feature-PCA representations, and the network's internal embedding represented the series through alternative coordinate systems that differed in both the information retained and the dependence among coordinates. ACF-Implied provides the clearest controlled case in which the fitted dynamic information was held fixed while its coordinate representation changed.

It would be natural to read this pattern through the classical account of finite-sample dimensionality, in which discrimination deteriorates as the number of coordinates grows relative to the information available (Hughes, 1968; Trunk, 1979). The present results locate the harm more specifically. Ambient dimension alone did not explain the pattern: in Study 1 the 32-dimensional CNN-Latent recovered memberships about as well as the one- or two-dimensional CNN-Pred (pooled means .204 and .212), and in Study 2 it had the highest pooled mean, whereas in the Unknown-K analysis the ten-dimensional ACF-Implied, which added no fitted information to the one- or two-dimensional Model-Based estimate, selected additional components in every no-subgroup replication.

A representation whose coordinates are deterministic functions of the fitted coefficients, as ACF-Implied's are, has the intrinsic dimension of those coefficients (one or two) inside whatever ambient dimension it occupies, and its coordinates are strongly dependent. Intrinsic dimension, covariance rank, and effective rank describe different aspects of that structure, and the present studies did not measure them separately. The Unknown-K contrast used full-covariance mixtures and varied the representation rather than the covariance specification, so it does not identify the mechanism. The practical consequence is that ambient dimension alone is insufficient to diagnose clustering difficulty. In the evaluated pipelines, coordinate dependence remains a candidate explanation for cluster-number calibration, while sample-size effects depended on how each representation was constructed and, for the deep learning representations, how the training sample was defined.

The Unknown-K analysis provided the clearest controlled demonstration of representation-dependent cluster-number selection in the present studies. ACF-Implied and Model-Based were derived from the same fitted autoregressive information. For AR(1), ACF-Implied expressed that information as a 10-dimensional model-implied autocorrelation vector, whereas Model-Based retained the original 1-dimensional coefficient. Under the null (no true subgroup structure, $K = 1$), ACF-Implied produced a false-positive rate of 100.0%, compared with 6.0% for Model-Based. Because the underlying fitted information was held constant across both representations, this difference cannot be attributed to estimation quality, training procedure, or alignment with the data-generating process. It is a property of how the same information was expressed as coordinates. The present analyses do not establish the mechanism linking coordinate structure to the selected component count. Representation choice is a structural determinant of clustering calibration, not only of condition-level ARI.

The CNN results are therefore not best interpreted as a simple failure of neural networks. Instead, they show that deep learning representations must be evaluated according to the downstream task. In Study 1 the network learned meaningful dynamic information, and clustering was somewhat more successful when that information was expressed as predicted coefficients than as the 32-dimensional embedding; in Study 2's low-information regime the embedding led, and both deep learning representations led direct estimation. The deep learning representations' value was conditional on the information available, not absent. This distinction is important because much of the deep-learning literature emphasizes predictive performance, whereas clustering requires the learned space to preserve between-person geometry in a way that supports unsupervised recovery (Ismail Fawaz et al., 2019; LeCun et al., 1998). For psychological time-series applications, this suggests that neural networks may be most useful when they produce interpretable low-dimensional summaries or when their latent spaces are explicitly trained for cluster separation rather than only for parameter prediction.

11.3 The Role of Series Length, Separation, and Observation Noise

Across both studies, subgroup recovery was strongly constrained by the amount and clarity of information available in the data. In Study 1, series length was the dominant factor: longer time series improved estimation precision and produced better subgroup recovery across methods. This finding is consistent with the basic intuition that autoregressive dynamics require adequate within-person observations to estimate reliably. Very short series provided too little temporal information for any method to recover subgroup structure well, regardless of representation. This has direct relevance for intensive longitudinal designs, where the number of observations per person is often limited by participant burden, compliance, and study duration (Hamaker & Wichers, 2017).

The dependence on series length also has a measurement interpretation. Wenzel and Brose (2023) report a variance-decomposition reliability estimate of approximately .52 for conventional two-step estimates of person-level emotional inertia, with reliability increasing substantially when unreliability is modeled directly. Although this estimate does not characterize the representations evaluated in the present simulations, it illustrates how measurement error in person-level dynamic estimates can obscure individual differences. Model-Based and ACF-Empirical similarly rely on person-level summaries estimated from finite series. Consistent with this concern, series length accounted for roughly half the variance in Model-Based recovery, whereas sample size accounted for almost none. For representations estimated separately for each individual, adding individuals does not directly improve the precision of each person's coordinates, whereas adding occasions provides more information for estimating them.

Beyond its effect on recovery accuracy, series length also changed the relative performance of the representations. In Study 1, longer series improved recovery across methods and widened the advantage of direct parameter estimation. In Study 2, however, increasing series length changed which representation led. At $T = 50$, the deep learning representations had the highest ARI in most conditions. At $T = 200$, Model-Based had the largest share among individual representations (42.6%), although the two deep learning representations together accounted for 57.4%. Thus, longer series increased the relative competitiveness of direct estimation without giving it a majority under the primary comparison rule. One plausible explanation is that deep learning representations can reduce the effects of noisy person-level estimates when temporal information is limited, whereas direct estimation becomes more competitive as additional observations improve coefficient precision. However, the present analyses did not independently decompose the bias and variance contributions to this pattern.

Observation noise, manipulated as signal-to-noise ratio, also played an important role. The manipulation is a specific operationalization of data quality and does not stand in for every form of psychological measurement reliability. Lower observation noise improved subgroup recovery, but its effects depended on the broader difficulty of the condition. When subgroup separation was favorable, improved signal quality sharpened the clustering structure. When separation was weak, however, reducing observation noise could not fully compensate for the lack of distinct subgroup dynamics. This point is practically important for psychological measurement. Intensive longitudinal data are often noisy, and increasing the number of observations cannot fully solve the problem if the measured process does not contain clear between-person dynamic structure. Good clustering therefore requires both adequate temporal sampling and low-noise measurement of the construct whose dynamics define the intended subgroups.

11.4 Fitting Strategy and the Construction of the Representation

The fitting-strategy finding reported in Section 8.1, that person-specific order selection before clustering produced a mixed-dimensional representation and near-chance recovery, has important implications for clustering workflows. In many applied settings, researchers may be tempted to estimate the "best" model separately for each individual and then cluster the resulting parameters. The present results suggest that such a strategy can be risky when selected models differ in dimensionality. Even if the model-selection procedure performs above chance, the resulting feature space may no longer reflect the grouping construct of interest. Instead, it may reflect the mechanics of model selection and feature padding. The relevant warning is therefore not that adaptive modeling is always inappropriate, but that adaptive feature construction must preserve a common and substantively meaningful geometry across individuals.

The persistence-space design of the primary Study 2 avoided the mixed-dimensional representation problem by fitting every individual with a common AR(2) specification, which contains AR(1) as the special case $\varphi_2 = 0$ at the latent-process level while requiring estimation of an additional coefficient for true AR(1) series. The broader lesson of both designs is the same: the fitted model should be chosen not only for individual-level adequacy but for whether the representation it yields has a common and substantively meaningful geometry across individuals. Where higher-order dynamics are substantively central, a common higher-order representation may be appropriate. Where subgroups are defined primarily by one-step persistence, however, estimating the extra coefficient can be costly, as the comparison of common AR(1) and AR(2) fitting in Section 8.1 suggests.

11.5 When Deep Learning Representations Helped

The deep learning representations earned their place in the low-information regime. In Study 2 they led on average, dominated the shortest series, and lost ground to direct estimation only as series lengthened. In Study 1 they had the highest ARI in 26 of the 30 conditions in which Model-Based did not, all at $N = 300$, where they had the most individuals to learn from. ACF-Empirical did in the other four. A plausible explanation is the bias–variance trade-off described in the Cross-Study Synthesis: OLS reproduces the sampling variability of each person's estimate, which blurs subgroup separation, whereas the network's predictions shrink toward the training distribution (standard-deviation ratios of 0.52 and 0.40 for the two coefficients in Study 2) and yield a more compact representation of the same individuals.

This interpretation should be treated as a mechanism suggested by the simulation results rather than as a definitive account of the CNNs' behavior. Both primary CNNs were trained on OLS-estimated coefficient targets, so the learned mapping could in principle regularize sampling

variability present in those targets rather than merely reproduce it. The Study 2 diagnostics are consistent with such contraction, but they do not isolate whether the effect arose from the neural architecture, the training distribution, the bounded output mapping, or another aspect of estimation. Future work could test this explanation directly by comparing OLS estimates, explicitly regularized estimators, hierarchical or Bayesian shrinkage estimates, and supervised predicted-parameter representations under matched generating conditions and targets.

11.6 Subgroup Recovery and Subgroup Enumeration Require Separate Evaluation

Recovering memberships when the number of clusters is supplied and determining how many clusters should be selected are distinct problems. A representation can recover memberships reasonably well at fixed K while frequently selecting additional components when no discrete subgroups exist. Conversely, a conservative selection procedure can control false positives while rarely recovering the correct count when discrete subgroups are present.

This distinction is especially important under continuous heterogeneity. A finite mixture may use several components to approximate a continuously varying population even when no discrete psychological types exist. The selected component count therefore reflects the mixture model preferred by the selection criterion for the evaluated representation. It does not establish the number of substantive populations.

Subgrouping analyses should consequently distinguish among membership recovery, correct-count selection, detection of any multicomponent solution, and evidence for substantively interpretable categorical structure. Evidence about one of these outcomes should not be treated as evidence about all of them.

11.7 Implications for Intensive Longitudinal Research

The present findings have several implications for researchers using intensive longitudinal data to identify latent subgroups. First, clustering should be guided by a clear theory of what distinguishes individuals dynamically. If subgroups are hypothesized to differ in inertia, persistence, or autoregressive feedback, then features that directly represent those dynamics may be preferable to broad feature sets that summarize many aspects of the series simultaneously. This is consistent with long-standing concerns in intensive longitudinal research that person-level dynamics should be modeled in ways that preserve within-person temporal structure rather than reducing repeated measures to static summaries (Hamaker & Wichers, 2017; Molenaar, 2004).

Second, researchers should be cautious when interpreting clusters recovered from redundant or strongly dependent feature spaces, regardless of dimension. Rich feature sets may obscure the dynamic signal defining subgroup membership. Internal clustering indices and visual separation do not guarantee meaningful dynamic subgroups when features contain noise, redundancy, or preprocessing artifacts. In the Unknown-K analysis, changing the coordinates of the same estimated information changed the selected component count without adding new person-level information, and in the empirical demonstration the selected number of clusters differed across representations of the same two affect series. Because dimensionality, coordinate dependence, and geometry changed together, the mechanism was not isolated. When external validation is limited, simulations help characterize how clustering pipelines behave under known conditions.

Third, model-selection decisions should be treated as part of the clustering pipeline rather than as neutral preprocessing. The fitting-strategy extension reported in Section 8.1, which is

distinct from the primary persistence-space Study 2, showed that strategy choice can change the entire recovery landscape. In applied work, this means that researchers should report not only the clustering algorithm, but also the model-fitting strategy, parameterization, feature-padding rules, scaling decisions, and any order-selection procedure used to construct the clustering input. Without those details, it is difficult to determine whether recovered clusters reflect psychological dynamics or artifacts of the feature-construction process.

Fourth, selected cluster counts should be interpreted alongside controls that clarify what selection means. In the empirical demonstration, BIC with the GMM selected two clusters for three representations, whereas the gap statistic with *k*-means selected one for every representation, and no null controls were run. In the separate Unknown-K analysis, with its own settings, CNN-Pred with BIC selected more than one component in 54.4% of no-subgroup data sets. More generally, matched null controls can reveal whether a selection procedure tends to favor additional components in the absence of discrete subgroups.

Li et al. (2025) reported three clusters with BIC and the Gaussian mixture under both temporal orders, and one with the gap statistic and *k*-means, for 227 participants from this project. The demonstration selected two and one for 178 participants defined by a different eligibility rule, so the two counts are not directly comparable and do not contradict each other. In both analyses, the cluster with more persistent negative affect also reported more depression. Neither count alone establishes substantive population structure.

Taken together, the simulations suggest practical guidance, with the caveat that no representation performed best in every condition. When individuals share a temporal order, Model-Based is a reasonable starting point: in Study 1 it had the highest ARI in 114 of 144 conditions, including all 72 with $N = 100$, and its advantage was clearest with long series and

well-separated subgroups; with larger samples and short series, CNN-Latent is also worth considering. When temporal order varies within subgroups, the deep learning representations and Model-Based performed similarly, with CNN-Latent more often highest for shorter series and Model-Based for longer series, and the differences between them were small. Across conditions, researchers should collect longer series where possible and not expect recovery from about 15 occasions, prefer model-based or deep learning representations over the feature-based representations evaluated here, fit one common temporal order and report it, and check the selected number of clusters against null controls.

11.8 Limitations

Several limitations bound the conclusions of this thesis. First, the simulations focused primarily on stationary univariate AR(1) and AR(2) processes. This controlled benchmark helped evaluate representation choice, temporal information, and fitting strategy, but psychological time series may also contain nonstationarity, nonlinear dynamics, regime changes, multivariate coupling, irregular spacing, missingness, and measurement reactivity. The findings therefore apply to the evaluated autoregressive settings rather than establishing a general ranking across intensive longitudinal data.

The empirical demonstration only partly relaxed this univariate focus. VAR(1), VAR(3), and the network used positive and negative affect together, but Model-Based, ACF-Empirical, and Feature-PCA summarized each affect separately and therefore omitted the cross-lagged relationships between them. Model-Based's selection of one cluster where VAR(1) and CNN-Pred selected two may partly reflect that omission.

Second, the primary subgroup structures were defined through autoregressive dynamics. This design favors representations that preserve the same information used to generate

membership. The scale-shift and volatility analyses demonstrate the resulting boundary clearly: when membership was instead defined by scale or conditional volatility, representations retaining those properties outperformed the autoregressive representations. The general conclusion is therefore not that parameter-based clustering is universally preferable, but that representation performance depends on alignment with the subgroup-defining signal.

Third, the analysis covered only a limited set of clustering and cluster-selection procedures. The primary recovery comparisons used diagonal-covariance Gaussian mixtures and *k*-means, while cluster-number analyses relied primarily on BIC or the gap statistic. Unlike the primary recovery comparisons, the Unknown-K analysis used full-covariance mixtures, as did the fitting-strategy extension for at least 126 of its 162 conditions. Shape-based approaches such as dynamic time warping and k-Shape were not evaluated. Other procedures, including hierarchical clustering, Bayesian approaches, and alternative selection criteria, may respond differently to representation dependence and dimensionality. Likewise, the BIC-Adaptive failure applies to the implemented zero-padding and routing procedure rather than to adaptive order selection in general.

Fourth, the neural results should not be generalized to all deep learning representations. Both primary simulations used one supervised convolutional architecture trained on OLS-estimated autoregressive coefficients from independently generated training series. The studies nevertheless differed in the other respects described in the Study 2 Method, so changes in relative performance cannot be attributed to population heterogeneity alone. In the empirical demonstration, the same architecture was trained on the participants it then represented, so its agreement with VAR(1) partly reflects its training target.

Finally, the empirical demonstration used one data set, single items for positive and negative affect, and 178 participants with nearly complete diaries, who may not represent the broader sample. About 30% of grid positions, including every overnight position, were imputed, so the results also rest on the imputation model. No null controls were run, the network was trained on the participants it represented, and one cluster's negative affect was near the floor of the scale, so its weak carry-over and the clusters' pre-diary differences may partly reflect level rather than dynamics. With six-item averages, the two-cluster solutions did not appear. The demonstration therefore shows how the pipelines behave in real data; it does not establish general subgroups.

11.9 Future Directions

Future work should extend the present benchmark in three main directions. First, the range of dynamic processes and subgroup targets should be broadened. Nonstationary, regime-switching, nonlinear, and multivariate processes may require representations that capture changing regimes, cross-variable relations, or other dynamics not summarized adequately by univariate autoregressive coefficients. Evaluating these settings would test whether the present findings about representation alignment and person-level information generalize to richer psychological processes.

Extending the evaluated representations and clustering procedures to multivariate series, including VAR-based representations, would incorporate within-variable and cross-variable lagged relationships in a common framework. Such representations could carry additional dynamic information, although whether they improve the resulting groupings, their stability, or their interpretability remains an empirical question. Multivariate models also require estimating

more parameters per person and therefore more within-person information. Their cross-lagged associations should not be interpreted causally.

Second, future studies should examine whether the advantages of deep learning representations in low-information settings reflect a broader benefit of regularization. Direct OLS estimates could be compared with penalized estimators, hierarchical or empirical Bayes shrinkage, and supervised parameter prediction under identical data-generating conditions. A deliberately controlled comparison of OLS-targeted and true-parameter-targeted networks could separately test the consequences of target noise, while holding architecture and population structure fixed. Such designs would clarify whether the observed contraction is a specifically neural effect or an instance of a more general shrinkage advantage.

Third, representation structure and cluster-number calibration warrant direct study. Adaptive strategies should be developed that preserve a common-dimensional representation without introducing artificial order axes, and the relationship among coordinate dependence, effective rank, covariance specification, and selected cluster count should be evaluated systematically across clustering algorithms. These analyses should be paired with additional empirical cohorts and calibrated null and positive controls to determine which diagnostics generalize beyond the present demonstration.

12 Conclusion

Across two Monte Carlo studies, robustness analyses and extensions, and an empirical demonstration, subgroup recovery depended on three closely related features of the analysis: what information the representation retained, how that information was organized for clustering, and how much temporal information was available for each individual. Direct autoregressive parameter estimates performed especially well when the subgroup-defining structure matched

those parameters and the individual series provided sufficient information. Deep learning representations were most competitive in lower-information settings, where their transformation of noisy person-level estimates could provide an advantage.

The results also show that membership recovery and subgroup enumeration cannot be treated as the same problem. Representations containing the same fitted dynamic information produced markedly different cluster-number behavior when that information was expressed through different coordinate structures, and in the empirical demonstration the selected number of clusters depended on both the representation and the selection procedure. Empirical subgroup claims therefore require attention not only to clustering performance but also to representation dependence, the clustering algorithm's covariance assumptions, and calibration against appropriate null and positive controls.

Most importantly, the findings do not identify a universally superior representation family. When subgroups were defined by autoregressive persistence, parameter-based and deep learning representations aligned with that signal performed best. When subgroups were defined by scale or volatility, representations retaining those properties were favored instead. The practical implication is to begin with a substantive hypothesis about what distinguishes individuals dynamically, construct a representation that preserves that information, and evaluate the resulting representation-and-algorithm pipeline under conditions that test both recovery and calibration.

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2403.07815

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery, 31*(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338

Beltz, A. M., Wright, A. G. C., Sprague, B. N., & Molenaar, P. C. M. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment, 23*(4), 447–458. https://doi.org/10.1177/1073191116648209

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLOS ONE, 8*(4), Article e60188. https://doi.org/10.1371/journal.pone.0060188

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). Springer.

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

Cheang, W.-K., & Reinsel, G. C. (2000). Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. *Journal of the American Statistical Association, 95*(452), 1173–1184. https://doi.org/10.1080/01621459.2000.10474318

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

Crawford, C. M., Park, J. J., Chow, S.-M., Ernst, A. F., Pipiras, V., & Fisher, Z. F. (2026). Penalized subgrouping of heterogeneous time series. *Multivariate Behavioral Research, 61*(4), 579–602. https://doi.org/10.1080/00273171.2026.2622120

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics, 23*(1), Article 205. https://doi.org/10.1186/s12859-022-04675-1

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., & Keogh, E. (2019). The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica, 6*(6), 1293–1305. https://doi.org/10.1109/JAS.2019.1911747

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

Ernst, A. F., Albers, C. J., & Timmerman, M. E. (2024). A comprehensive model framework for between-individual differences in longitudinal data. *Psychological Methods, 29*(4), 748–766. https://doi.org/10.1037/met0000585

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

Ernst, A. F., Timmerman, M. E., Ji, F., Jeronimus, B. F., & Albers, C. J. (2024). Mixture multilevel vector-autoregressive modeling. *Psychological Methods, 29*(1), 137–154. https://doi.org/10.1037/met0000551

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Fisher, Z. F., Kim, Y., Fredrickson, B. L., & Pipiras, V. (2022). Penalized estimation and forecasting of multiple subject intensive longitudinal data. *Psychometrika, 87*(2), 403–431. https://doi.org/10.1007/s11336-021-09825-7

Franceschi, J.-Y., Dieuleveut, A., & Jaggi, M. (2019). Unsupervised scalable representation learning for multivariate time series. *Advances in Neural Information Processing Systems, 32*. https://arxiv.org/abs/1901.10738

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Fulcher, B. D., & Jones, N. S. (2017). hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems, 5*(5), 527–531.e3. https://doi.org/10.1016/j.cels.2017.10.001

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), Article 20130048. https://doi.org/10.1098/rsif.2013.0048

Galeano, P., & Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo, 4*(4), 383–403.

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*(1), 310–319. https://doi.org/10.1016/j.neuroimage.2012.06.026

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning* (PMLR Vol. 235, pp. 16115–16152). https://proceedings.mlr.press/v235/goswami24a.html

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

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van

Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine, 42*(5), 903–920. https://doi.org/10.1017/S0033291711001966

Henderson, T., & Fulcher, B. D. (2021). An empirical evaluation of time-series feature sets. In *2021 International Conference on Data Mining Workshops (ICDMW)* (pp. 1032–1038). IEEE. https://doi.org/10.1109/ICDMW53433.2021.00134

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters, 64*, 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hennig, C., & Lin, C.-J. (2015). Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing, 25*(4), 821–833. https://doi.org/10.1007/s11222-015-9566-5

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218. https://doi.org/10.1007/BF01908075

Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55–63. https://doi.org/10.1109/TIT.1968.1054102

Hyndman, R. J., Wang, E., & Laptev, N. (2015). Large-scale unusual time series detection. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1616–1619). IEEE. https://doi.org/10.1109/ICDMW.2015.104

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery, 33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A, 374*(2065), Article 20150202. https://doi.org/10.1098/rsta.2015.0202

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting, 33*(2), 345–358. https://doi.org/10.1016/j.ijforecast.2016.09.004

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics, 80*(2), 218–230. https://doi.org/10.1162/003465398557465

Kreienkamp, J., Agostini, M., Monden, R., Epstude, K., de Jonge, P., & Bringmann, L. F. (2025). A gentle introduction and application of feature-based clustering with psychological time series. *Multivariate Behavioral Research, 60*(2), 362–392. https://doi.org/10.1080/00273171.2024.2432918

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97. https://doi.org/10.1002/nav.3800020109

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

Lafabregue, B., Weber, J., Gançarski, P., & Forestier, G. (2022). End-to-end deep representation learning for time series clustering: A comparative study. *Data Mining and Knowledge Discovery, 36*(1), 29–81. https://doi.org/10.1007/s10618-021-00796-y

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing, 13*(2), 163–167. https://doi.org/10.1023/A:1023260610025

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, Y., Song, H., & Jeronimus, B. (2025). Impact of temporal order selection on clustering intensive longitudinal data based on vector autoregressive models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000747

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

Liu, S., Gates, K. M., & Ferrer, E. (2024). Homogeneity assumptions in the analysis of dynamic processes. *Multivariate Behavioral Research, 59*(6), 1166–1176. https://doi.org/10.1080/00273171.2023.2225172

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., & Jones, N. S. (2019). catch22: CAnonical time-series characteristics. *Data Mining and Knowledge Discovery, 33*(6), 1821–1852. https://doi.org/10.1007/s10618-019-00647-x

Madiraju, N. S., Sadat, S. M., Fisher, D., & Karimabadi, H. (2018). *Deep temporal clustering: Fully unsupervised learning of time-domain features* [Preprint]. arXiv. https://arxiv.org/abs/1802.01059

Maharaj, E. A. (2000). Cluster of time series. *Journal of Classification, 17*(2), 297–314. https://doi.org/10.1007/s003570000023

Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: A review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery, 38*(4), 1958–2031. https://doi.org/10.1007/s10618-024-01022-1

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21*(4), 441–458. https://doi.org/10.1207/s15327906mbr2104_5

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

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

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology, 6*, 109–138. https://doi.org/10.1146/annurev.clinpsy.121208.131413

Paparrizos, J., & Bogireddy, S. P. T. R. (2025). Time-series clustering: A comprehensive study of data mining, machine learning, and deep learning methods. *Proceedings of the VLDB Endowment, 18*(11), 4380–4395. https://doi.org/10.14778/3749646.3749700

Paparrizos, J., Yang, F., & Li, H. (2024). *Bridging the gap: A decade review of time-series clustering methods* [Preprint]. arXiv. https://arxiv.org/abs/2412.20582

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*, 8024–8035. https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis, 11*(2), 153–164. https://doi.org/10.1111/j.1467-9892.1990.tb00048.x

Ren, B., & Barnett, I. (2022). Autoregressive mixture models for clustering time series. *Journal of Time Series Analysis, 43*(6), 918–937. https://doi.org/10.1111/jtsa.12644

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96). https://doi.org/10.25080/Majora-92bf1922-011

Shader, T. M., & Beauchaine, T. P. (2022). A Monte Carlo evaluation of growth mixture modeling. *Development and Psychopathology, 34*(4), 1604–1617. https://doi.org/10.1017/S0954579420002230

Shaman, P., & Stine, R. A. (1988). The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association, 83*(403), 842–848. https://doi.org/10.1080/01621459.1988.10478672

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

Shu, Y., Levi, P. T., Tiego, J., Constable, T., Pang, J. C., Hartshorn, B., Kwee, J., Fortune, K., Bellgrove, M., & Fornito, A. (2026). Complex affect dynamics offer limited incremental value for cross-sectional prediction of psychological and behavioural variables in cross-

validated models. *BMC Medicine*. Advance online publication. https://doi.org/10.1186/s12916-026-04936-3

Song, H., & Ferrer, E. (2012). Bayesian estimation of random coefficient dynamic factor models. *Multivariate Behavioral Research, 47*(1), 26–60. https://doi.org/10.1080/00273171.2012.640593

Song, H., & Zhang, Z. (2014). Analyzing multiple multivariate time series data using multilevel dynamic factor models. *Multivariate Behavioral Research, 49*(1), 67–77. https://doi.org/10.1080/00273171.2013.851018

Steinley, D. (2004). Properties of the Hubert–Arabie adjusted Rand index. *Psychological Methods, 9*(3), 386–396. https://doi.org/10.1037/1082-989X.9.3.386

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B, 63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*(3), 306–307. https://doi.org/10.1109/TPAMI.1979.4766926

Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2022). Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining and Knowledge Discovery, 12*(3), Article e1444. https://doi.org/10.1002/widm.1444

van der Krieke, L., Blaauw, F. J., Emerencia, A. C., Schenk, H. M., Slaets, J. P. J., Bos, E. H., de Jonge, P., & Jeronimus, B. F. (2017). Temporal dynamics of health and well-being: A

crowdsourcing approach to momentary assessments and automated generation of personalized feedback. *Psychosomatic Medicine, 79*(2), 213–223. https://doi.org/10.1097/PSY.0000000000000378

van der Krieke, L., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B. K., Emerencia, A. C., Schenk, H. M., de Vos, S., Snippe, E., Wichers, M., Wigman, J. T. W., Bos, E. H., Wardenaar, K. J., & de Jonge, P. (2016). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International Journal of Methods in Psychiatric Research, 25*(2), 123–144. https://doi.org/10.1002/mpr.1495

Van Mechelen, I., Hennig, C., & Kiers, H. A. L. (2024). Onset of a conceptual outline map to get a hold on the jungle of cluster analysis. *WIREs Data Mining and Knowledge Discovery, 14*(5), Article e1547. https://doi.org/10.1002/widm.1547

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

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 1578–1585). IEEE. https://doi.org/10.1109/IJCNN.2017.7966039

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, 23*(2), 412–424. https://doi.org/10.1037/emo0001108

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83. https://doi.org/10.2307/3001968

Xiong, Y., & Yeung, D.-Y. (2004). Time series clustering with ARMA mixtures. *Pattern Recognition, 37*(8), 1675–1689. https://doi.org/10.1016/j.patcog.2003.12.018

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., & Xu, B. (2022). TS2Vec: Towards universal representation of time series. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(8), 8980–8987. https://doi.org/10.1609/aaai.v36i8.20881

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2114–2124). ACM. https://doi.org/10.1145/3447548.3467401

Table 1

Study 1 Subgroup Recovery by Representation Across 144 Conditions

Representation	*M*	MCSE	Pooled-best clustering algorithm	GMM only	*k*-means only	Highest ARI	Share
Model-Based	.290	.001	*k*-means	.280	.290	114	79.2%
CNN-Pred	.212	<.001	GMM	.212	.203	4	2.8%
CNN-Latent	.204	<.001	GMM	.204	.179	22	15.3%
ACF-Empirical	.165	<.001	*k*-means	.153	.165	4	2.8%
Feature-PCA	.153	<.001	GMM	.153	.133	0	0.0%

Note. M is the pooled mean adjusted Rand index across the 144 $K = 3$ clustering conditions, each representation under its pooled-best clustering algorithm. Highest ARI uses the within-condition best-algorithm rule; under this rule Model-Based had the highest ARI in 114 of 144 (79.2%), compared with 110 of 144 (76.4%) under the diagonal Gaussian mixture alone and 134 of 144 (93.1%) under *k*-means alone. Single-algorithm means are reported to show that the ranking does not depend on the assignment. Head-to-head counts use each representation under its globally pooled-best clustering algorithm, the same basis as the inferential tests: Model-Based leads CNN-Pred in 142 conditions, ACF-Empirical in 139, CNN-Latent in 119, and Feature-PCA in all 144. Outright and pairwise counts differ because a third representation can have the highest ARI in a cell.

Table 2

Study 1 ANOVA Factor Hierarchy for Model-Based ARI

Rank	Factor / interaction	η^2	Interpretation
1	Series length (*T*)	50.7%	Largest share
2	Separation	31.4%	Second-largest share
3	Signal-to-noise ratio	7.4%	Moderate observation-noise effect
4	*T* × Separation	6.7%	*T* matters most when separation is favorable
5	AR order	0.3%	Modest AR(2) penalty
6	Sample size (*N*)	0.1%	Minimal for Model-Based

Note. Values are from a descriptive factorial ANOVA on Model-Based condition-level mean ARI under the diagonal Gaussian mixture, across the 144 $K = 3$ clustering conditions. The design is balanced, so one-way and Type II decompositions coincide. Under *k*-means the same hierarchy holds with *T* at 53.2% and separation at 30.4%.

Table 3

Cluster-Number Selection Across 48 Null Conditions and 144 Three-Subgroup Conditions

Representation	Selection procedure	Null false-positive rate	Correct selection of $K = 3$
Model-Based	GMM-BIC	6.0%	3.9%
CNN-Latent	Gap statistic	8.1%	8.9%
ACF-Empirical	GMM-BIC	18.4%	2.7%
CNN-Pred	GMM-BIC	54.4%	38.8%
Feature-PCA	GMM-BIC	64.6%	12.2%
ACF-Implied	GMM-BIC	100.0%	21.7%

Note. The false-positive rate is the proportion of simulation replications selecting more than one component under the 48 null conditions. Correct-count selection is the proportion selecting $K = 3$ under the 144 three-subgroup conditions. Each condition contained 500 replications. The selection procedure is identified separately because CNN-Latent used the gap statistic in the primary summary, whereas the other representations used GMM-BIC. ACF-Implied carries the same fitted coefficient information as Model-Based, expressed in ten deterministic coordinates rather than one or two. CNN-Pred showed a sample-size dependence not visible in the pooled rate, rising from 12.5% at $N = 100$ to 96.2% at $N = 300$, whereas Model-Based moved from 2.6% to 9.4%. Under the gap statistic, CNN-Pred false-positive selection falls to 14.7% and correct-count selection to 4.1%. No representation combined a low false-positive rate with a high correct-count rate.

Table 4

Study 2 Subgroup Recovery and Distribution of the Highest ARI Across 162 Conditions

Representation	GMM *M*	*k*-means *M*	Highest ARI	Share	% at *T* = 50	% at *T* = 100	% at *T* = 200
CNN-Latent	.2515	.2352	75	46.3%	53.7	51.9	33.3
CNN-Pred	.2458	.2417	40	24.7%	25.9	24.1	24.1
Model-Based	.2391	.2384	47	29.0%	20.4	24.1	42.6
ACF-Empirical	.1560	.1712	0	0.0%	0.0	0.0	0.0
Feature-PCA	.0507	.0743	0	0.0%	0.0	0.0	0.0

Note. Pooled mean adjusted Rand index under each clustering algorithm across the 162 persistence-space conditions. Highest ARI uses the within-condition best-algorithm rule. The three rightmost columns give each representation's share of the 54 conditions at each series length, the underlying counts being 29, 28, and 18 for CNN-Latent, 14, 13, and 13 for CNN-Pred, and 11, 13, and 23 for Model-Based out of 54 conditions at each length; the rise in the Model-Based share and the fall in the CNN-Latent share across *T* is the information-regime pattern discussed in the text. Monte Carlo standard errors of the pooled means were .0002 to .0004.

Table 5

Study 2 ANOVA Factor Hierarchy for Model-Based ARI (Persistence-Space Design)

Rank	Factor / interaction	η^2	Interpretation
1	Series length (*T*)	43.5%	Largest share, closely followed by separation
2	Separation	40.6%	Nearly equal to series length
3	Signal-to-noise ratio	12.1%	Third, well below the first two
4	Order composition	0.9%	Minimal for Model-Based
5	*T* × Separation	0.8%	Largest interaction, but small
6	Sample size (*N*)	0.6%	Minimal for Model-Based
7	Separation × SNR	0.3%	Negligible
8	*T* × Composition	0.1%	Negligible

Note. Values are from a descriptive Type II ANOVA on Model-Based condition-level mean ARI under the diagonal Gaussian mixture across the 162 persistence-space conditions of Study 2. Residual = 1.0%.

Table 6

Fitting-Strategy Extension: AR1-Fit ANOVA Factor Hierarchy for Model-Based ARI

Rank	Factor / interaction	η^2	Interpretation
1	Separation	58.9%	Dominant under heterogeneity
2	Series length (*T*)	22.7%	Important, but secondary
3	Signal-to-noise ratio	14.8%	Third, after separation and series length
4	*T* × Separation	1.6%	*T* benefits depend on separation
5	Separation × SNR	1.2%	SNR benefits depend on separation
6	Sample size (*N*)	0.2%	Minimal for Model-Based
7	Order composition	0.2%	Minimal for Model-Based

Note. η^2 values are from a descriptive Type II ANOVA on condition-level mean ARI under AR1-fit across the 162 conditions of the fitting-strategy extension reported in Section 8.1. This table belongs to that extension, not to the primary persistence-space Study 2.

Table 7

Number of Clusters Selected for Each Representation

Representation	Values per participant	GMM (BIC) *K*	ΔBIC	Cluster sizes	Gap statistic *K*
VAR models					
VAR(1)	4	2	24.3	93, 85	1
VAR(3)	12	2	40.6	89, 89	1
Five representations					
Model-Based	2	1	8.7	178	1
CNN-Pred	4	2	19.3	98, 80	1
CNN-Latent	27	7[a]	—	—	1
ACF-Empirical	20	1	17.8	178	1
Feature-PCA	2	1	0.2	178	1

Note. $N = 178$; 1 to 8 clusters compared. GMM (BIC) = Bayesian information criterion with the Gaussian mixture model, which also chose the covariance structure; ΔBIC = lead of the selected number of clusters over the next-best number; cluster sizes are for the GMM solution; gap statistic = gap statistic with *k*-means. Model-Based is one AR(1) coefficient per affect, without the cross-lagged coefficients of VAR(1).

[a] BIC kept improving because many latent values were exactly zero; the count is not interpretable.

Table 8

Characteristics of the Two VAR(1) Clusters

Characteristic	Cluster 1	Cluster 2
n (%)	93 (52.2)	85 (47.8)
Age, *M* (*SD*)	40.1 (12.9)	45.1 (15.4)
Female, %	82.8	81.2
Momentary PA, *M* (*SD*)	56.9 (20.3)	63.3 (17.8)
Momentary NA, *M* (*SD*)	22.2 (22.0)	6.0 (9.8)
Correlation of PA and NA	−.52***	−.19

Note. Clusters from VAR(1) with the GMM (Table 7). Momentary PA and NA are items 5 and 8 (0–100) over all completed occasions. Correlation = correlation of participant-mean PA with participant-mean NA within the cluster. PA = positive affect; NA = negative affect.

***$p < .001$.

Table 9

Cluster-Specific VAR(3) Coefficients

Cluster	Outcome	PA $t-1$	NA $t-1$	PA $t-2$	NA $t-2$	PA $t-3$	NA $t-3$
Cluster 1	PA t	0.23	−0.10	−0.02	0.00	0.03	−0.01
	NA t	−0.02	0.36	0.00	0.00	−0.01	0.05
Cluster 2	PA t	0.26	0.06	−0.02	0.00	0.05	0.02
	NA t	0.00	0.10	0.00	0.01	−0.01	0.02

Note. Cluster means of participant-specific VAR(3) coefficients (Li et al., 2025); clusters as in Table 8. Rows are outcomes at time t; columns are predictors one to three steps earlier (one step = 6 hours). PA = positive affect; NA = negative affect.

Table 10

Pre-Diary Measures by Cluster

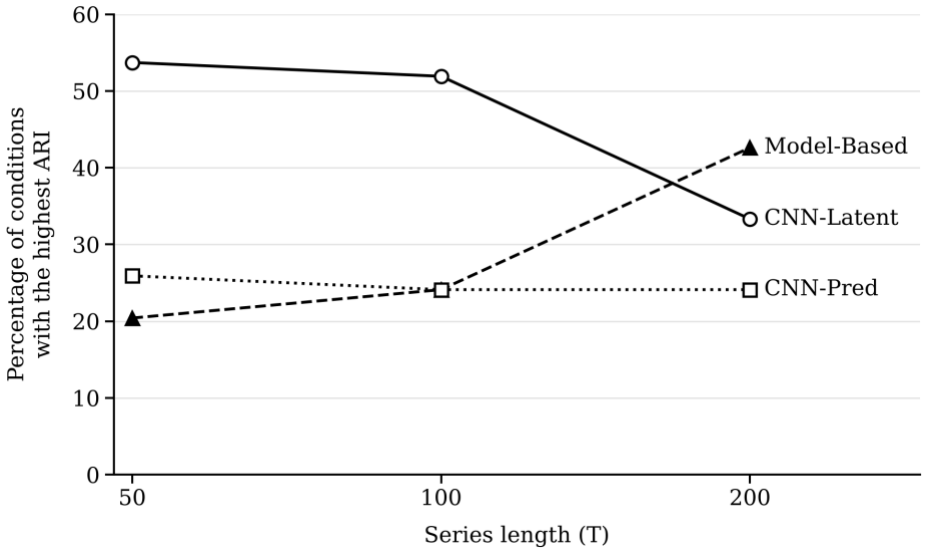
Measure	Cluster 1 *M* (*SD*)	Cluster 2 *M* (*SD*)	*F*	*p*
Depression (QIDS)	8.20 (5.00)	4.76 (3.43)	28.09	< .001
Depression (DASS-21)[a]	12.58 (4.34)	9.67 (2.68)	16.25	< .001
Anxiety/panic	1.19 (0.89)	0.51 (0.67)	33.67	< .001
Positive affect	30.58 (7.38)	34.09 (6.26)	11.63	< .001
Negative affect	23.70 (7.37)	17.12 (4.69)	49.46	< .001

Note. One-way analyses of variance, $df = 1, 176$; all five differences remain significant after Holm correction. Clusters as in Table 8. QIDS = Quick Inventory of Depressive Symptomatology; DASS-21 = Depression Anxiety Stress Scales, depression scale.

[a] $n = 103$ ($df = 1, 101$).

Figure 1

Share of Study 2 Conditions in Which Each Leading Representation Had the Highest ARI Across Series Lengths



Note. Each point is the percentage of the 54 conditions at that series length in which the representation achieved the highest mean adjusted Rand index, using the within-condition best-algorithm rule. ACF-Empirical and Feature-PCA did not have the highest ARI in any condition at any series length and are omitted. The crossing of the Model-Based and CNN-Latent series between $T = 100$ and $T = 200$ is the information-regime pattern described in the text; the pooled means in Table 4 do not show it.