Received: 29 November 2019 | Accepted: 30 April 2021

DOI: 10.1111/rssa.12712

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Clustering longitudinal life-course sequences using mixtures of exponential-distance models

Keefe Murphy[1] | **T. Brendan Murphy**[2,3] | **Raffaella Piccarreta**[4] | **I. Claire Gormley**[2,3]

[1]Department of Mathematics and Statistics, Maynooth University, Maynooth, Ireland

[2]School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

[3]Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

[4]Department of Decision Sciences, Università Bocconi, Milano, Italy

Correspondence
Keefe Murphy, Department of Mathematics and Statistics, Maynooth University, Maynooth, Ireland.
Email: keefe.murphy@mu.ie

Funding information
Science Foundation Ireland, Grant/Award Number: SFI/12/RC/2289_P2; MIUR-PRIN, Grant/Award Number: 20177BRJXS

Abstract

Sequence analysis is an increasingly popular approach for analysing life courses represented by ordered collections of activities experienced by subjects over time. Here, we analyse a survey data set containing information on the career trajectories of a cohort of Northern Irish youths tracked between the ages of 16 and 22. We propose a novel, model-based clustering approach suited to the analysis of such data from a holistic perspective, with the aims of estimating the number of typical career trajectories, identifying the relevant features of these patterns, and assessing the extent to which such patterns are shaped by background characteristics. Several criteria exist for measuring pairwise dissimilarities among categorical sequences. Typically, dissimilarity matrices are employed as input to heuristic clustering algorithms. The family of methods we develop instead clusters sequences directly using mixtures of exponential-distance models. Basing the models on weighted variants of the Hamming distance metric permits closed-form expressions for parameter estimation. Simultaneously allowing the component membership probabilities to depend on fixed covariates and accommodating sampling weights in the clustering process yields new insights on the Northern Irish data. In

particular, we find that school examination performance is the single most important predictor of cluster membership.

KEYWORDS

exponential-distance models, gating covariates, life-course sequences, model-based clustering, survey sampling weights, weighted Hamming distance

1 | INTRODUCTION

Sequence analysis (SA) is an umbrella term for tools defined to explore and describe categorical life-course data. Specifically, attention is focused on the ordered sequence of states (or activities) experienced by individuals over a given time span (usually at T equally spaced discrete time periods). Here we focus on the transition from school to work for a cohort of Northern Irish youths, using survey data obtained from the 1999 sweep of the Status Zero Survey (McVicar, 2000; McVicar & Anyadike-Danes, 2002)—henceforth referred to as the MVAD data—in which, for each individual, a sequence of monthly labour market activities experienced between the ages of 16 and 22 is recorded.

Typically, the goal of SA is to identify the most relevant patterns in the data. To this end, pairwise dissimilarities among sequences in their entirety are first assessed. Dissimilarity matrices are then employed to identify the most typical trajectories using, in the vast majority of applications, cluster analysis. These problems are receiving increasing attention in the demographic and social literature, also due to the increasing number of retrospective as well as prospective longitudinal studies, such as the British Household Panel Survey (BHPS)[1] and the subsequent larger and more wide-ranging UK Household Longitudinal Study (Understanding Society)[2], or the Socio-Economic Panel for Germany (SOEP)[3], the National Longitudinal Surveys for the USA (NLS)[4] and the Generations & Gender Programme for selected European countries (GGP)[5]. All of these surveys, much like the MVAD data considered in this paper, collect information about labour market activities, as well as other significant life events.

Quantifying the distance between such categorical sequences is not a trivial task. Optimal matching (OM), developed by Abbott and Forrest (1986) and extended to sociology by Abbott and Hrycak (1990), is popular among the SA community. OM is derived from the edit distance originally proposed in the field of information theory and computer science by Levenshtein (1966). The OM metric assigns costs to the different types of edits, namely insertion, deletion, and substitution. Typically, insertion and deletion are assigned a cost of 1 while substitution costs are allowed to vary. However, specifying these costs often involves subjective choices, which may lead to violations of the triangle inequality if not done carefully. Several proposals in the literature introduced criteria to improve or guide the choice of costs in OM; Muñoz-Bullón and Malo (2003), for instance estimate the substitution-cost matrix in a data-driven fashion using the between-states transition rates. Alternative dissimilarity criteria have also been introduced to allow control over the importance assigned to the characteristics

[1]https://www.iser.essex.ac.ul.bhps.

[2]https://www.understandingsociety.ac.uk.

[3]https://www.diw.de/en/soep.

[4]https://www.bls.gov/nls/.

[5]https://www.ggp-i.org.

of the sequences (namely, the collection of experienced states, their timing or their duration) in the assessment of their differences: see Studer and Ritschard (2016) for an excellent discussion. Even so, there are no results proving that one procedure is superior to others and the choice of dissimilarity measure remains a fundamental question for researchers.

Given a dissimilarity matrix $\mathbf{D}$ obtained from a set of sequences $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$, where n is the number of subjects, cluster analysis is usually applied to group sequences and identify the most typical trajectories experienced by the sampled individuals. Heuristic clustering algorithms, either hierarchical or partitional, are typically used. In many applications, it is also of interest to relate sequences to a set of baseline covariates. Within the described framework, this is solely done by relating the uncovered hard clustering partition to covariates using, for example, multinomial logistic regression (MLR). This approach was adopted in McVicar and Anyadike-Danes (2002), after applying Ward's agglomerative clustering algorithm (Ward, 1963) to an OM dissimilarity matrix to obtain $G = 5$ clusters of the MVAD sequences, without performing model selection. Such an approach is questionable from a few points of view. First, the original sequences are substituted by a categorical variable indicating cluster membership, thus disregarding the heterogeneity within clusters. This is clearly only sensible when the clusters are sufficiently homogeneous, otherwise sequences which are weakly related to clusters would be regarded as similar to those in their cluster. However, a clear clustering structure can often be obtained only by increasing the number of clusters (often with some clusters possibly small in size). More importantly, suitable partitions do not necessarily lead to suitable response variables as input for the MLR. It thus seems desirable to cluster sequences and relate the clusters to the covariates simultaneously.

Thus, the aim of our analysis is threefold: to estimate the number of typical trajectories in the MVAD data, to identify the relevant features of these patterns and to establish to what extent such patterns are shaped by the individuals' background characteristics, as captured by a set of baseline covariates measured at age 16. To address these issues, we propose to cluster the MVAD sequences in a model-based fashion, allowing the covariates to affect the soft cluster membership probabilities, rather than leaving them exogenous to the clustering model. This permits us to better understand if and to what extent the typical sequence patterns characterising each cluster are affected by specific covariates. Model-based clustering methods typically assume that the data arise from a finite mixture of G distributions; Bouveyron et al. (2019) provide an excellent overview. In principle, any distribution(s) can be used, though the term 'model-based clustering' was popularised by Banfield and Raftery (1993), in which the component distributions are assumed to be parsimoniously parameterised multivariate Gaussians with component-specific parameters. Such models have been recently extended to the mixture of experts setting (Gormley & Frühwirth-Schnatter, 2019) to facilitate dependence on fixed covariates (Murphy & Murphy, 2020). However, these models can be problematic when applied to dissimilarity matrices, either due to non-identifiability or because the input data are usually far from Gaussian. This problem cannot be addressed by applying multidimensional scaling to $\mathbf{D}$ because the resulting low-dimensional configuration is also typically far from Gaussian. Notably, our attempts to fit non-Gaussian mixtures in these settings did not yield useful results.

Another popular framework for clustering categorical data is latent class analysis (LCA; Lazarsfeld & Henry, 1968). Agresti (2002) shows the connection between model-based clustering and LCA. Such models are finite mixtures in which the component distributions are assumed to be multi-way cross-classification tables with all variables mutually independent. Latent class regression models (LCR; Dayton & Macready, 1988) are particularly interesting, because their connection to the mixture of experts framework permits the inclusion of covariates to predict the latent class memberships. However, fitting such models is challenging when the sequence length, the number of categories, or the number of latent classes are even moderately large, due to the explosion in the number of parameters.

Evidently, there is a conflict of perspectives between the model-based and the heuristic, distance-based approaches to clustering in the SA community. For this reason, and the others mentioned above, we model the sequences directly (in the sense that the sequences themselves are treated as inputs, rather than **D**) with the implicit substitution costs which define the distance metric being estimable parameters of a generative probability model rather than inputs (either estimated or subjectively specified), via **D**, to a heuristic clustering algorithm. This is achieved using parsimonious mixtures of exponential-distance models, which typically depend on a central sequence and a precision parameter in a way that relates to the chosen distance metric. Our framework for analysing the MVAD data, as a model-based approach which nonetheless relies on distances, thus reconciles the aforementioned conflict.

Mostly for reasons of computational convenience, we use dissimilarities based on simple matching, in particular the Hamming distance (Hamming, 1950). Although the focus on substitution operations has the sociological advantage of targeting trajectories with contemporaneous similarities—in contrast to the prohibited insertion and deletion operations, which focus on matching states irrespective of their timing—this distance is liable to suffer from temporal rigidity, since anticipations and/or postponements of the same choices in life courses are not accounted for. Hence, similar sequences shifted by one time period may be maximally distant from one another. While misalignment is less of a concern for sequences exhibiting long durations in the same state, we address the issue using weighted variants of the Hamming distance, characterised by a range of constraints on the precision parameters in the mixture setting. This leads to the novel MEDseq model family, which can be seen as similar to a version of the *k*-medoids/PAM algorithm (Kaufman & Rousseeuw, 1990, Chapter 2) based on the Hamming distance with some restrictions relaxed. We defer the comparison to Section 4.2 as the parallels relate to technical issues of model estimation.

Importantly, information is also available with the MVAD data on the survey sampling weights, which are only incorporated in the MLR stage of the analysis in McVicar and Anyadike-Danes (2002). While sampling weights can be incorporated into heuristic clustering algorithms, such as Ward's hierarchical clustering (by weighting the linkages between clusters) or *k*-medoids, and subsequently in the MLR, one of the advantages of our approach is that both the covariates and the weights are incorporated simultaneously. This is achieved by leveraging the model-based paradigm; the weights are incorporated by appropriately weighting the likelihood function and the covariates are incorporated by assuming they influence the soft component membership probabilities.

MEDseq models, like standard SA heuristic clustering algorithms and LCA models, approach the clustering task from the holistic perspective of treating trajectories as *whole* units of analysis, in order to uncover groups of similar sequences. In contrast, a number of multistate models employing finite mixtures with Markov components (e.g. Melnykov, 2016a; Pamminger & Frühwirth-Schnatter, 2010) or with hidden Markov components (Helske et al., 2016) have recently attained popularity for the analysis of categorical sequence data. Such models focus on modelling instantaneous transitions within the life course and on factors that might explain the probability of experiencing them. As described by Wu (2000), this amounts to a difference between considering sequences in their entirety under the MEDseq framework or as time-to-event processes under the Markovian framework. Indeed, as our aim is to establish sequence typologies for the MVAD data, a holistic approach is preferable to Markovian approaches. The former concentrates on questions of global similarities and considers the full richness of the trajectories without discarding the details of episode ordering, duration, or transition (Muñoz-Bullón & Malo, 2003), while the latter framework makes the often unsuitable simplifying assumption that trajectories can be efficiently summarised only by their recent past (Piccarreta & Studer, 2019).

The remainder of the article is organised as follows. Section 2 presents some exploratory analysis of the MVAD data. Section 3 develops the MEDseq family of mixtures of exponential-distance

models. Section 4 describes the model fitting procedure and discusses factors affecting performance. Section 5 presents results for the MVAD data, including applications of MEDseq models and comparisons to other methods. The insights gleaned from the MVAD data under the optimal MEDseq model are summarised in Section 6. The paper concludes with a discussion on the MEDseq methodology and potential future extensions in Section 7. A software implementation of the full MEDseq model family is provided by the associated R package `MEDseq` (Murphy et al., 2020). The package was developed specifically for this application and is available from https://www.r-project.org (R Core Team, 2020).

2 | STATUS ZERO SURVEY: MVAD DATA

The term 'MVAD data' refers throughout to a cohort of $n = 712$ Northern Irish youths aged 16 and eligible to leave compulsory education as of July 1993 who were observed at monthly intervals until June 1999 as part of the Status Zero Survey (Armstrong et al., 1997; McVicar, 2000; McVicar & Anyadike-Danes, 2002). The subjects were interviewed about the labour market activities they experienced, distinguishing between employment (EM), further education (FE), higher education (HE), joblessness (JL), school (SC) or training (TR). Each observation i is represented by an ordered categorical sequence of length $T = 72$, with an alphabet $\mathcal{V}$ of size $v = 6$ possible states, for example $\mathbf{s}_i = \left(s_{i,1}, s_{i,2}, \ldots, s_{i,72}\right)^\top = (\text{SC}, \text{SC}, \ldots, \text{TR}, \text{TR}, \ldots, \text{EM}, \text{EM})^\top$. The sequences share a common length, the time periods are equally spaced and there are no missing data. In the context of the Northern Irish education system at the time, SC refers to secondary school, which may be a grammar school to which entrance is granted upon completion of an exam. At age 16, students take General Certificate of Secondary Education (GCSE) examinations; students who do well are eligible to continue in school for a further 2 years (to sit A-level exams) or to leave, for example to a training/apprenticeship programme (TR). Further education (FE) is distinguished from higher education (HE); FE typically refers to applied post-GCSE courses while HE refers to third-level/university courses, typically pursued at age 18 after the successful completion of A-level exams. Notably, the transitions HE $\rightsquigarrow$ SC and TR $\rightsquigarrow$ HE are never observed.

It is of interest to relate the MVAD sequences to covariates in order to understand whether different characteristics (related to gender, community, geographic and social conditions, and personal abilities) impact on the school-to-work trajectories. These covariates are summarised in Table 1. All covariates were measured at the age of 16 (i.e. at the start of the study period in July 1993), with the exception of 'Funemp' and 'Livboth', and are thus static background characteristics. As achieving 5 or more grades at A–C in GCSE exams is the traditional cut-off point for progression to the additional two-years of secondary school required for a transition to HE, we expect the 'GCSE5eq' covariate in particular to be strongly associated with the clustering. The MVAD data also come with associated observation-specific survey sampling weights, which depend on the 'Grammar' and 'Location' covariates. Specifically, the sample was stratified in such a way that a predetermined number of subjects were in each state, for each location and both school types, immediately after the end of the compulsory education period (Armstrong et al., 1997).

The MVAD data are available in the R packages `MEDseq` and `TraMineR` (Gabadinho et al., 2011). As the data have been used to illustrate some of the functionalities of the `TraMineR` package in its associated vignette[6], interesting features of an exploratory analysis of the data can be found therein. However, we reproduce plots of the transversal state distributions in Figure 1 and the transversal entropies in Figure 2, that is the Shannon entropies of the state distributions at each time point

[6]https://cran.r-project.org/web/packages/TraMineR/vignettes/TraMineR-state-sequence.pdf.

TABLE 1 Available covariates for the MVAD data set. For binary covariates, the event denoted by 1 is indicated. Otherwise, the levels of the categorical covariate 'Location' are grouped in curly brackets

Covariate	Description
Catholic	1 = yes
FMPR	SOC code of father's current or most recent job as of the beginning of the survey, 1 = SOC1 (Standard Occupational Classification: professional, managerial, or related)
Funemp	Father's employment status as of June 1999, 1 = employed
GCSE5eq	Qualifications gained by the end of compulsory education, 1 = 5+ GCSE grades at A–C, or equivalent
Gender	1 = male
Grammar	Type of secondary education, 1 = grammar school
Livboth	Living arrangements as of June 1995, 1 = living with both parents
Location	Location of school, one of five Education and Library Board areas in Northern Ireland, {Belfast, North Eastern, South Eastern, Southern, Western}

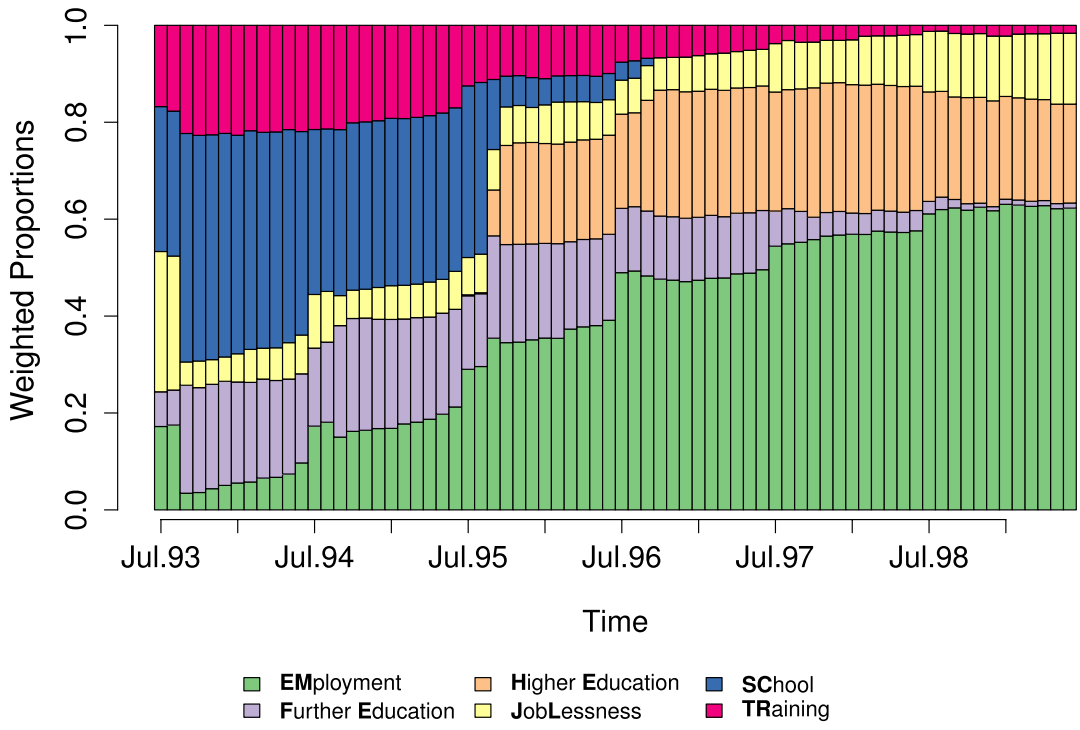

FIGURE 1 Overall state distribution for the weighted MVAD data, coloured by state [Colour figure can be viewed at wileyonlinelibrary.com]

(Billari, 2001), with the sampling weights accounted for in both cases. Notably, fewer than v states are observed in certain months.

Figure 1 shows that the number of subjects who found employment increased over time. Conversely, fewer students were in training or further education by the end of the observation period. Most students appear to have entirely left school within 2/3 years of the commencement of the survey. Finally, while students only reached the age of 18 and began to pursue higher education from July 1995 onwards, a number of students had already pursued further education during the two preceding years.

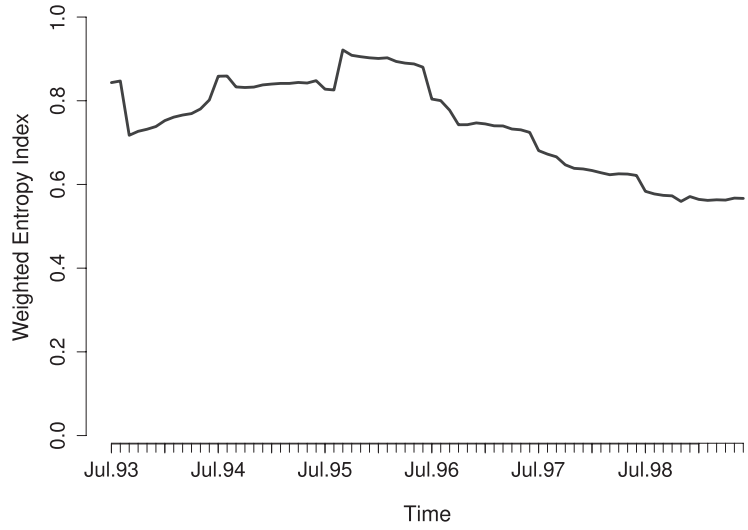

FIGURE 2 Transversal entropy plot for the weighted MVAD data

Figure 2 confirms that the level of heterogeneity in the state distribution varies over time. In particular, the entropy declines after Sep 1995, by which point most students had left school.

Interestingly, many students were jobless during the first two months of observation. As the vast majority of cases notably remained in the same state in this period, which coincided with the summer break from school, all subsequent analyses are conducted on a version of the data with the first two time points removed. Hence, we work hereafter with sequences of length $T = 70$, commencing with the return to school in September 1993. As the sampling design depends on 'Grammar' and 'Location', the term 'all covariates' henceforth refers to all *other* covariates in Table 1. While Murphy and Murphy (2020) show that the same covariate can affect more than one part of a mixture of experts model, and in different ways, removing the quantities used to define the weights eases the interpretability of the results.

3 | MODELLING

In this section, we introduce the novel family of MEDseq models. The exponential-distance model is described in Section 3.1, extended to account for the sampling weights in Section 3.2, expanded into a family of mixtures in Section 3.3, and finally, embedded within the mixture of experts framework in Section 3.4 in order to accommodate the available covariates.

3.1 | Exponential-distance models

For an arbitrary distance metric $\mathrm{d}(\cdot, \cdot)$, location parameter $\boldsymbol{\theta}$, and precision parameter λ, the probability mass function (PMF) of an exponential-distance model (EDM) for sequences is

$$f\left(\mathbf{s}_i|\boldsymbol{\theta},\, \lambda,\, \mathrm{d}\right) = \frac{\exp\left(-\lambda \mathrm{d}\left(\mathbf{s}_i,\, \boldsymbol{\theta}\right)\right)}{\sum_{\boldsymbol{\sigma}\in S_v^T} \exp\left(-\lambda\, \mathrm{d}\left(\boldsymbol{\sigma}, \boldsymbol{\theta}\right)\right)} = \Psi\left(\lambda, \boldsymbol{\theta}|T, v\right)^{-1} \exp\left(-\lambda\, \mathrm{d}\left(\mathbf{s}_i, \boldsymbol{\theta}\right)\right), \tag{1}$$

with the corresponding log-likelihood function given by

$$\ell(\boldsymbol{\theta}, \lambda|\mathbf{S}, \mathrm{d}) = \sum_{i=1}^{n} \log f\left(\mathbf{s}_i|\boldsymbol{\theta}, \lambda, \mathrm{d}\right) = -\lambda \sum_{i=1}^{n} \mathrm{d}\left(\mathbf{s}_i, \boldsymbol{\theta}\right) - n\log\Psi(\lambda, \boldsymbol{\theta}|T, v). \quad (2)$$

Such a model is analogous to the Gaussian distribution (characterised by the squared Euclidean distance from the mean) and similar to the Mallows model for permutations (Mallows, 1957). Indeed, mixtures of Mallows models have been used to cluster rankings (Murphy & Martin, 2003). We only consider models with $\lambda \geq 0$. When $\lambda = 0$, the distribution of sequences is uniform. For $\lambda > 0$, the central sequence $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_T\right)$ is the mode, that is the sequence with highest probability, and the probability of any other sequence decays exponentially as its distance from $\boldsymbol{\theta}$ increases. The precision parameter λ controls the speed of this decay. Larger λ values cause sequences to concentrate around $\boldsymbol{\theta}$, tending towards a point-mass as $\lambda \to \infty$. Notably, λ is not identifiable when all sequences are identical.

The log-likelihood in Equation (2) is generally intractable, as the normalising constant $\Psi(\lambda, \boldsymbol{\theta}|T, v)$ depends on the parameter λ (under OM and other, more complicated distances, it can also depend on $\boldsymbol{\theta}$), as well as the fixed constants $T > 1$ and $v > 1$, and requires a sum over all possible sequences. With reference to the MVAD data, for example, computing $\Psi(\lambda,\boldsymbol{\theta}|T,v)$ is practically infeasible as there are $\mathrm{card}\left(\mathcal{S}_v^T\right) = v^T = 6^{70}$ possible sequences. Fortunately, however, the normalising constant exists in closed form under the Hamming distance, $\mathrm{d_H}\left(\mathbf{s}_i, \mathbf{s}_j\right) = \sum_{t=1}^{T} \mathbb{1}\left(s_{i,t} \neq s_{j,t}\right)$, in a manner which facilitates direct enumeration and crucially does not depend on $\boldsymbol{\theta}$, as a sum with only $T+1$ terms. Consider, for example, the Hamming distances between all ternary ($v = 3$) sequences of length $T = 4$. From the arbitrary reference sequence (0, 0, 0, 0), there is 1 count of a distance of 0, 8 counts of a distance of 1, 24 counts of a distance of 2, 32 counts of a distance of 3, and 16 counts of a distance of 4. Thus, $\Psi_{\mathrm{H}}(\lambda|T = 4, v = 3) = e^0 + 8e^{-\lambda} + 24e^{-2\lambda} + 32e^{-3\lambda} + 16e^{-4\lambda}$. Hence, the normalising constant under the Hamming distance metric depends on the parameter λ, the sequence length T, and the number of categories v, and simplifies greatly:

$$\Psi_{\mathrm{H}}(\lambda|T, v) = \sum_{p=0}^{T} \binom{T}{p} (v-1)^p \exp(-\lambda p) = \left((v-1)e^{-\lambda} + 1\right)^T. \quad (3)$$

Inspired by the generalised Mallows model (Irurozki et al., 2019), the EDM in Equation (1) based on the Hamming distance can be extended to one based on the weighted Hamming distance. By introducing T precision parameters $\lambda_1, \ldots, \lambda_T$, one for each time point (i.e. sequence position), and expressing the exponent in Equation (1) as $\mathrm{d_{WH}}(\mathbf{s}_i, \boldsymbol{\theta}\,|\,\lambda_1, \ldots, \lambda_T) = \sum_{t=1}^{T} \lambda_t \mathbb{1}\left(s_{i,t} \neq \theta_t\right)$ rather than $\lambda \mathrm{d_H}(\mathbf{s}_i, \boldsymbol{\theta}) = \lambda \sum_{t=1}^{T} \mathbb{1}\left(s_{i,t} \neq \theta_t\right)$, different time periods can contribute differently to the overall distance, weighted according to the period-specific precision parameters. Thus, the distance from $\boldsymbol{\theta}$ to $\mathbf{s}_i$ under $\mathrm{d}_{WH}(\cdot, \cdot\,|\cdot)$ becomes a sum of the λ_t values associated with each time point which differs from the corresponding θ_t, across the whole sequence. This acts as implicit variable selection and allows modelling situations in which there is high consensus regarding the state values of some time periods, with large uncertainty about the values of others. Accounting for the alignment of contemporaneous matchings in this way helps to prevent sequences with the same (Hamming) distance from $\boldsymbol{\theta}$ from having the same probability. Given that sequences equidistant from $\boldsymbol{\theta}$ can nevertheless exhibit element-wise mismatches between themselves, this may help later, in the mixture setting, to induce stronger between-cluster separation and within-cluster homogeneity. The non-constant transversal entropies in Figure 2 suggest that this extension may also be fruitful in terms of capturing different degrees of dispersion in the state distributions of the MVAD data over time. Crucially, the various benefits outlined above can be achieved without any

tractability sacrifices. The log-likelihood in Equation (2) is merely rewritten with the weighted Hamming distance decomposed into its T components and the normalising constant in Equation (3) also modified accordingly:

$$\ell\left(\boldsymbol{\theta}, \lambda_1, \ldots, \lambda_T | \mathbf{S}, \mathrm{d}_{\mathrm{WH}}\right) = -\sum_{i=1}^{n}\left[\sum_{t=1}^{T}\left(\lambda_t \mathbb{1}\left(s_{i,t} \neq \theta_t\right) + \log\left((v-1)\, e^{-\lambda_t} + 1\right)\right)\right].$$

Though other dissimilarity measures are available for sequences, we henceforth consider measures based on the Hamming distance only, chiefly for the computational reasons outlined above. In our setting, $\lambda \mathrm{d}_{\mathrm{H}}(\cdot, \cdot)$ can be seen as a special case of OM with all substitution costs set to λ and no insertions or deletions. As it has time-varying substitution costs, $\mathrm{d}_{\mathrm{WH}}(\cdot, \cdot | \cdot)$ is similar to the dynamic Hamming distance (Lesnard, 2010), a prominent alternative to OM. However, such costs in our models are always assumed to be common with respect to each pair of states. Hence, $\mathrm{d}_{\mathrm{WH}}(\cdot, \cdot | \cdot)$ equates to the Gower distance between nominal variables (Gower, 1971) with equally weighted states and unequally weighted time points.

3.2 | Incorporating sampling weights

Sampling weights are often associated with life-course data, as the data typically arise from surveys where the weights are used to correct for representivity bias under stratified sampling designs. Following Chambers and Skinner (2003), the sampling weights $\mathbf{w} = \left(w_1, \ldots, w_n\right)$ are incorporated into the EDM by exponentiating the likelihood of each sampled unit by the attached weight w_i, which is akin to unit i being observed w_i times. The resultant pseudo likelihood $\mathcal{L}^{\mathbf{w}}(\cdot | \cdot)$ reweights the likelihood contribution for each unit in order to rebalance the information in the observed sample to approximate the balance of information in the target finite population. The sampling weights $\mathbf{w}$ are thus interpretable as being inversely proportional to the unit inclusion probabilities, remain fixed, and are confined to those included in the sample. Notably, $f\left(\mathbf{s}_i | \boldsymbol{\theta}, \lambda, \mathrm{d}\right)^{w_i} \propto f\left(\mathbf{s}_i | \boldsymbol{\theta}, w_i \lambda, \mathrm{d}\right)$, such that the weights induce a unit-specific rescaling of the precision parameter; it follows that the observed data are independent but not identically distributed.

A secondary benefit of incorporating weights is that it facilitates computational gains in the presence of duplicate cases. Such duplicates are likely when dealing with discrete life-course data. This non-uniqueness can be exploited using likelihood weights for computational efficiency, by fitting models to the subset of unique sequences only, weighted by the sum of the sampling weights (if available, otherwise $w_i = 1 \forall\, i$) across each corresponding set of duplicates. In modifying $\mathbf{w}$ in this way, cases with different sampling weights which are otherwise duplicates are also treated as duplicates, in such a way that the (pseudo) likelihood is unaltered. The number of duplicates clearly lowers when considering both the sequences themselves and their associated covariate patterns. In particular, all cases are unique when there are continuous covariates. Nonetheless, in the MVAD data, and in many applications, the covariates are all categorical. Hence, exploiting non-uniqueness in this manner can be extremely computationally convenient. For instance, only 490 of the $n = 712$ sequences in the MVAD data set are distinct. However, to avoid notational confusion, all subsequent expressions are written as though duplicate cases have not been discarded.

Though the weights for the MVAD data sum to ≈ 711.52, we henceforth follow Xu et al. (2013) in always assuming that the weights have been normalised to sum to the sample size n. In doing so, subsequent expressions are simplified further and the use of model selection criteria (see Section 4.3) relying on the pseudo likelihood is facilitated. While the resultant rescaling of the MVAD weights is negligible, we note that multiplying $\mathbf{w}$ by a scalar does not affect parameter estimation.

3.3 | A family of mixtures of exponential-distance models

Extending the EDM based on the Hamming distance with sampling weights to the model-based clustering setting yields a pseudo likelihood function of the form

$$\mathcal{L}^{\mathbf{w}}\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \lambda | \mathbf{S}, \mathbf{w}, \mathrm{d}_{\mathrm{H}}\right) = \prod_{i=1}^{n}\left[\sum_{g=1}^{G}\tau_g \frac{\exp\left(-\lambda \mathrm{d}_{\mathrm{H}}\left(\mathbf{s}_i, \boldsymbol{\theta}_g\right)\right)}{\left((v-1)e^{-\lambda}+1\right)^{T}}\right]^{w_i},$$

where the mixing proportions $\tau_1, \ldots, \tau_G$ are positive and sum to 1. Thus, the clustering approach is both model-based and distance-based, thereby bridging the gap between these two 'cultures' in the SA community.

The mixture setting naturally suggests a further extension, whereby the precision parameter λ can be constrained or unconstrained across clusters, in addition to the aforementioned allowance for the precision parameters to vary (or not) across time points. Within a family of models, we term 'MEDseq', we thus define the CC, UC, CU and UU models, where the first letter denotes whether precision parameters are constrained (C) or unconstrained (U) across clusters and the second denotes the same across time points. Notably, all models deviate from the simple matching distance on which they are based, as even the most constrained CC model could be said to employ a weighted variant thereof, by virtue of allowing for $\lambda \neq 1$. The model family allows moving between more parsimonious models and more heavily parameterised, flexible models which may provide a better fit to the data. As the precision parameters relate to the substitution costs characterising variants of the Hamming distance, quantities used to define the overall distance measure are allowed to vary in different ways, while still being treated as model parameters rather than inputs. In particular, models with names beginning with U reflect scenarios in which the implicit substitution costs differ across clusters. Hence, the UU model is analogous to the hierarchical Ward$_p$ algorithm (de Amorim, 2015), in the sense of having cluster-specific feature weights (albeit with no tuning required).

Given the role played by λ when it takes the value 0, whereby the state distribution is uniform, it is convenient and natural to include a noise component (denoted by N), whose single precision parameter is fixed to 0, to robustify inference by capturing deviant cases and minimising their deleterious effects on parameter estimation for the other, more defined clusters. Adding this extension to each of the 4 models above, regardless of how precision parameters are otherwise specified, completes the MEDseq model family with the CCN, UCN, CUN, and UUN models. When $G = 1$, the CC, CU, and CCN models can be fitted. When $G = 2$, the UCN and UUN models are equivalent to the CCN and CUN models, respectively, as there is only one non-noise component. As the noise component arises naturally from restricting the parameter space, we consider the noise component as one of the G components, denoted hereafter with the subscript 0. All 8 model types are summarised further in Appendix A.

3.4 | Incorporating covariates

We now illustrate how to incorporate the available covariate information into the clustering process, both to guide the construction of the clusters and to find the typical trajectories which can be best predicted by covariates. As is typical for model-based clustering analyses, the data are augmented in MEDseq models by introducing a latent cluster membership indicator vector $\mathbf{z}_i = \left(z_{i,1}, \ldots, z_{i,G}\right)^{\top}$, where $z_{i,g} = 1$ if observation i belongs to cluster g and $z_{i,g} = 0$ otherwise. The MEDseq approach can be easily extended to incorporate the possible effects of covariates on the assignments of sequences

to clusters by allowing the covariates to influence the distribution of the latent variable $\mathbf{z}_i$. Thus, such covariates are interpreted differently from those used to define the sampling weights, if any.

The inclusion of covariates is achieved under the mixture of experts framework (Gormley & Frühwirth-Schnatter, 2019; Jacobs et al., 1991), by extending the mixture model to allow the mixing proportions for observation i to depend on covariates $\mathbf{x}_i$. This, rather than having covariates enter through the component distributions, is particularly attractive, as the interpretation of the remaining component-specific parameters is the same as it would be under a model without covariates. For example, in the case of the CC MEDseq model

$$f\left(\mathbf{s}_i|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_G,\lambda,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_G,\mathbf{x}_i,\mathrm{d_H}\right)^{w_i}=\left[\sum_{g=1}^{G}\tau_g\left(\mathbf{x}_i|\boldsymbol{\beta}_g\right)\frac{\exp\left(-\lambda\mathrm{d_H}\left(\mathbf{s}_i,\boldsymbol{\theta}_g\right)\right)}{\left((v-1)\,e^{-\lambda}+1\right)^{T}}\right]^{w_i},$$

where the mixing proportions $\tau_g\left(\mathbf{x}_i|\boldsymbol{\beta}_g\right)$ (henceforth $\tau_g\left(\mathbf{x}_i\right)$, for simplicity) are referred to as the 'gating network', with $\tau_g\left(\mathbf{x}_i\right) > 0$ and $\sum_{g=1}^{G}\tau_g\left(\mathbf{x}_i\right) = 1$, as usual, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ are the gating network regression parameters. Such a model can be seen as a conditional mixture model (Bishop, 2006) because, given the covariates $\mathbf{x}_i$, the distribution of the sequences is a finite mixture model under which $\mathbf{z}_i$ has a multinomial distribution with a single trial and probabilities equal to $\tau_g\left(\mathbf{x}_i\right)$. The distance-based k-medoids algorithm, though closely related (see Section 4.2), does not accommodate the inclusion of gating covariates in this way.

Incorporating covariates in 'hard' clustering algorithms using MLR, as done by McVicar and Anyadike-Danes (2002), has been criticised because the hard assignment of extraneous cases can negatively impact internal cluster cohesion and the MLR coefficient estimates (Piccarreta & Studer, 2019). An advantage of the noise component in MEDseq models is that it captures uniformly distributed sequences that deviate from those in the other, more defined clusters. Filtering outliers in this way lessens their impact on the non-noise gating network coefficients, thereby enabling more accurate inference and improving the interpretability of the effects of the covariates. Moreover, the 'soft' partition obtained under the model-based paradigm allows the cluster membership probabilities for sequences lying on the boundary between two neighbouring clusters to be quantified and the effect of such sequences on the gating network coefficients to be mitigated.

As per Murphy and Murphy (2020), the CCN, UCN, CUN and UUN models which include an explicit noise component can be restricted to having covariates only influence the mixing proportions for the non-noise components, with all observations therefore assumed to have equal probability of belonging to the uniform noise component (i.e. by replacing $\tau_0\left(\mathbf{x}_i\right)$ with τ_0). We refer to the former setting as the gated noise (GN) setting and to the latter as the non-gated noise (NGN) setting. The NGN setting is the more parsimonious one, makes more clear the distinction between EDM components and the uniform component, and is particularly apt when τ_0 is expected to be small and/or the sequences are expected to overwhelm the gating covariate(s) in determining which cases are noise. Gating covariates can only be included when $G \geq 2$ under the GN setting or when there are 2 or more non-noise components under the NGN setting.

4 | MODEL ESTIMATION

This section describes our model-fitting approach and some implementation issues that arise in practice. Specifically, Section 4.1 outlines the ECM algorithm employed for parameter estimation, Section 4.2 discusses the initialisation thereof with reference to the similarities between MEDseq models and

the k-medoids and k-modes (Huang, 1997) algorithms, and the issues of model selection, covariate selection, and model validation are treated in Section 4.3.

4.1 | Model fitting via ECM

Parameter estimation is greatly simplified by the existence of a closed-form expression for the normalising constant for MEDseq models based on the Hamming or weighted Hamming distances. We focus on maximum (pseudo) likelihood estimation using a simple variant of the EM algorithm (Dempster et al., 1977). For simplicity, model fitting details are described chiefly for the CC MEDseq model with sampling weights and gating covariates. Additional details for other model types are deferred to Appendix B; so, too, are technical details pertaining to estimation of the precision parameter(s). The complete data pseudo likelihood for the CC model is given by

$$\mathcal{L}_c^{\mathbf{w}}\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \lambda, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G | \mathbf{S}, \mathbf{X}, \mathbf{Z}, \mathbf{w}, \mathrm{d_H}\right) = \prod_{i=1}^{n}\left[\prod_{g=1}^{G}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\lambda \mathrm{d_H}\left(\mathbf{s}_i, \boldsymbol{\theta}_g\right)\right)}{\left((v-1)\,e^{-\lambda}+1\right)^T}\right)^{z_{i,g}}\right]^{w_i},$$

and the complete data pseudo log-likelihood hence has the form:

$$\begin{aligned}\ell_c^{\mathbf{w}}\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \lambda, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G | \mathbf{S}, \mathbf{X}, \mathbf{Z}, \mathbf{w}, \mathrm{d_H}\right) = &\sum_{i=1}^{n}\sum_{g=1}^{G} z_{i,g} w_i [\log \tau_g(\mathbf{x}_i) - \lambda \mathrm{d_H}(\mathbf{s}_i, \boldsymbol{\theta}_g) \\ &- T\log((v-1)e^{-\lambda}+1)].\end{aligned} \tag{4}$$

Under this model, the distribution of $\mathbf{s}_i$ depends on the latent cluster membership variable $\mathbf{z}_i$, which in turn depends on covariates $\mathbf{x}_i$, while $\mathbf{s}_i$ is independent of $\mathbf{x}_i$ conditional on $\mathbf{z}_i$.

The iterative algorithm for MEDseq models follows in a similar manner to that for standard mixture models. It consists of an E-step (expectation) which replaces for each observation the missing data $\mathbf{z}_i$ with their expected values $\widehat{\mathbf{z}}_i$, which sum to 1, followed by a M-step (maximisation), which maximises the expected complete data pseudo log-likelihood. The M-step consists of a series of conditional maximisation (CM) steps in which each parameter is maximised individually, conditional on the other parameters remaining fixed. Hence, model fitting is in fact conducted using an expectation conditional maximisation (ECM) algorithm (Meng & Rubin, 1993). Aitken's acceleration criterion is used to assess convergence of the non-decreasing sequence of weighted pseudo log-likelihood estimates (Böhning et al., 1994). Parameter estimates produced on convergence achieve at least a local maximum of the pseudo likelihood function. Upon convergence, cluster memberships are estimated via the maximum a posteriori (MAP) classification, that is cases are assigned to the cluster g to which they most probably belong via $\mathrm{MAP}\left(\widehat{\mathbf{z}}_i\right) = \arg\max_{g\in\{1,\ldots,G\}}(\widehat{z}_{i,g})$.

The E-step (with similar expressions when λ is unconstrained across clusters and/or time points) involves computing the expression in Equation (5), where $(m+1)$ is the current iteration number:

$$\widehat{z}_{i,g}^{(m+1)} = \mathbb{E}\left(z_{i,g} | \mathbf{s}_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_g^{(m)}, \widehat{\lambda}^{(m)}, \widehat{\boldsymbol{\beta}}_g^{(m)}, w_i, \mathrm{d_H}\right) = \frac{\widehat{\tau}_g^{(m)}(\mathbf{x}_i) f(\mathbf{s}_i | \widehat{\boldsymbol{\theta}}_g^{(m)}, \widehat{\lambda}^{(m)}, \mathrm{d_H})}{\sum_{h=1}^{G}\widehat{\tau}_h^{(m)}(\mathbf{x}_i) f(\mathbf{s}_i | \widehat{\boldsymbol{\theta}}_h^{(m)}, \widehat{\lambda}^{(m)}, \mathrm{d_H})}. \tag{5}$$

Note that the weights w_i appear in neither the numerator nor the denominator, leaving the E-step unchanged regardless of the inclusion or exclusion of weights.

Subsequent subsections describe the CM-steps for estimating the remaining parameters in the model. These individual CM-steps rely on the current estimates $\widehat{\mathbf{Z}}^{(m+1)} = (\widehat{\mathbf{z}}_1^{(m+1)}, \ldots, \widehat{\mathbf{z}}_n^{(m+1)})$ to provide estimates of the gating network regression coefficients $\widehat{\boldsymbol{\beta}}_g^{(m+1)}$, and hence the mixing proportion parameters $\widehat{\tau}_g^{(m+1)}(\mathbf{x}_i)$, as well as the central sequence(s) $\widehat{\boldsymbol{\theta}}_g^{(m+1)}$ and component precision parameter(s) $\widehat{\lambda}^{(m+1)}$, though technical details for the latter, as they are the element which distinguishes the various MEDseq model types, are deferred to Appendix B. It is clear from Equation (4) that the sampling weights can be accounted for by simply multiplying every $\widehat{\mathbf{z}}_i^{(m+1)}$ by the corresponding weight w_i. Conversely, in the CM-steps which follow, corresponding formulas for unweighted MEDseq models can be recovered by replacing $\widehat{z}_{i,g}^{(m+1)} w_i$ with $\widehat{z}_{i,g}^{(m+1)}$.

4.1.1 | Estimating the gating network coefficients

The portion of Equation (4) corresponding to the gating network, given by $\sum_{i=1}^{n}\sum_{g=1}^{G} z_{i,g} w_i \log \tau_g(\mathbf{x}_i)$, is of the same form as a MLR model with weights given by w_i, here written with component 1 as the baseline reference level for identifiability reasons:

$$\log\frac{\tau_g(\mathbf{x}_i)}{\tau_1(\mathbf{x}_i)} = \log\frac{\Pr(z_{i,g}=1)}{\Pr(z_{i,1}=1)} = \tilde{\mathbf{x}}_i\boldsymbol{\beta}_g \,\forall\, g \geq 2, \text{ with } \boldsymbol{\beta}_1 = (0, \ldots, 0)^{\top},$$

where $\tilde{\mathbf{x}}_i = (1, \mathbf{x}_i)$. Thus, methods for fitting such models, with $\widehat{\mathbf{Z}}^{(m+1)}$ as the response, can be used to estimate the gating network regression parameters $\widehat{\boldsymbol{\beta}}_g^{(m+1)}$. As closed-form updates are unavailable for MLR coefficients, due to the nonlinear numerical optimisation involved, this step merely increases (rather than maximises) the expectation of this term. However, the monotonicity of the sequence of pseudo log-likelihood estimates is preserved and convergence is still guaranteed. Subsequently, the mixing proportions are given by

$$\widehat{\tau}_g^{(m+1)}(\mathbf{x}_i) = \frac{\exp(\tilde{\mathbf{x}}_i\widehat{\boldsymbol{\beta}}_g^{(m+1)})}{\sum_{h=1}^{G}\exp(\tilde{\mathbf{x}}_i\widehat{\boldsymbol{\beta}}_h^{(m+1)})}.$$

Conversely, $\boldsymbol{\tau}$ is estimated exactly via $\widehat{\tau}_g^{(m+1)} = n^{-1}\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i$ when there are no gating covariates. Since $\sum_{i=1}^{n} w_i = n$, this is simply the weighted mean of the gth column of the matrix $\widehat{\mathbf{Z}}^{(m+1)}$. However, $\boldsymbol{\tau}$ can also be constrained to be equal (i.e. $\tau_g = 1/G \,\forall\, g$) across clusters. Thus, situations where $\tau_{i,g} = \tau_g(\mathbf{x}_i)$, $\tau_{i,g} = \tau_g$, or $\tau_{i,g} = 1/G$ are accommodated.

The standard errors of the gating network's MLR at convergence are not a valid means of assessing the uncertainty of the coefficient estimates as the cluster membership probabilities are estimated rather than fixed and known. Therefore, we adapt the weighted likelihood bootstrap (WLBS) of O'Hagan et al. (2019) to the MEDseq setting. This is implemented by multiplying the sampling weights $\mathbf{w}$ by draws from an n-dimensional symmetric uniform Dirichlet distribution and refitting the MEDseq model. To ensure stable estimation of the standard errors, $B = 1000$ such samples are used here. To ensure rapid convergence and to circumvent label-switching problems, the estimated $\widehat{\mathbf{Z}}$ matrix from the original model is used to initialise the ECM algorithm for each sample with new likelihood weights. Finally, the standard errors of the gating network coefficients across the B samples are obtained. Although this approach does not produce fully valid variance estimates when there are sampling weights which arise from stratified designs, we adopt the WLBS in what follows in order to provide approximate standard errors. This issue is particularly pronounced when the probability of being included in the sample depends

on quantities being modelled. This concern provides additional justification for the aforementioned removal of the Grammar and Location covariates from our analysis.

4.1.2 | Estimating the central sequences

The location parameter $\boldsymbol{\theta}$ is sometimes referred to as the Fréchet mean or the central sequence. The k-medoids/PAM algorithm, which is closely related to the MEDseq models with certain restrictions imposed (see Section 4.2), fixes the estimate of $\widehat{\boldsymbol{\theta}}_g$ to be the medoid of cluster g (Kaufman & Rousseeuw, 1990), that is the observed sequence $\mathbf{s}_i \in \mathbf{S}$ with minimum weighted distance from the others currently assigned to the same cluster. This estimation approach is especially quick as the Hamming distance matrix for the observed sequences is pre-computed. Notably, this greedy search strategy may fail to find the optimum solution.

However, for a $G = 1$ unweighted EDM based on the Hamming distance, the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is given simply by the modal sequence, meaning that each $\widehat{\theta}_t$ is independently given by the most frequent state at the tth time point. This is intuitive when $\mathrm{d_H}\left(\mathbf{s}_i, \boldsymbol{\theta}\right)$ is expressed as $T - \sum_{t=1}^{T} \mathbb{1}\left(s_{i,t} = \theta_t\right)$, as $\widehat{\boldsymbol{\theta}}$ maximises the number of element-wise agreements. Thus, the parameter has a natural interpretation. For more complicated distance metrics, the first-improvement algorithm (Hoos & Stützle, 2004) or a genetic algorithm could be used to estimate $\boldsymbol{\theta}$. Notably, the modal sequence need not be an observed sequence in $\mathbf{S}$. It is also notable that any $\widehat{\theta}_t$ may be non-unique under any of the proposed estimation strategies. Such ties, if any, are broken at random.

For $G > 1$, under the ECM framework, central sequence position estimates $\widehat{\theta}_{g,t}^{(m+1)}$ are given by $\arg\max_{\vartheta \in \mathcal{V}_t}\left(\sum_{i=1}^{n} \widehat{z}_{i,g}^{(m+1)} w_i \mathbb{1}\left(s_{i,t} = \vartheta\right)\right)$, where $\mathcal{V}_t$ is the subset of $v_t \leq v$ states observed at time point t across all cases. As this expression is independent of the precision parameter(s), it holds for all MEDseq model types, including those based on weighted Hamming distance variants. Thus, $\widehat{\boldsymbol{\theta}}_g^{(m+1)}$ is similarly estimated easily and exactly via a weighted mode (much like k-modes), whereby each $\widehat{\theta}_{g,t}^{(m+1)}$ is given by the category corresponding to the maximum of the sum of the weights $\widehat{z}_{i,g}^{(m+1)} w_i$ associated with each of the v_t observed state values. Similarly, the central sequence under a weighted $G = 1$ model is also estimated via a weighted mode, with the weights given only by w_i. Notably, to estimate the central sequences for a MEDseq model of any type without sampling weights, one need only remove w_i from these terms. Note also that $\boldsymbol{\theta}_0$ does not need to be estimated for models with an explicit noise component as it does not contribute to the likelihood.

4.2 | ECM initialisation and comparison to k-medoids

MEDseq models share relevant features with the PAM algorithm. Both consider sequences from a holistic perspective and both rely on distances to a cluster centroid. However, PAM treats the matrix of pairwise distances between sequences as a pre-computed input, while under MEDseq models the distances to the centroids (and the costs which define the distance metric) are recomputed at each iteration, with the sequences themselves as input. Otherwise, compared to PAM based on the Hamming distance, MEDseq models differ only in that (i) $\boldsymbol{\theta}_g$ is estimated by the modal sequence rather than the medoid, (ii) $\boldsymbol{\tau}$ is estimated, or dependent on covariates via $\tau_g\left(\mathbf{x}_i\right)$, rather than constrained to be equal, (iii) λ is free to vary across clusters and/or time points, rather than being implicitly set to 1, (iv) a noise component can be included, and (v) the ECM algorithm rather than the classification EM algorithm (CEM; Celeux & Govaert, 1992) is used. The CEM algorithm employed by PAM uses *hard*

assignments $\tilde{z}_{i,g}$, computed in its C-step, such that $\tilde{z}_{i,g}^{(m+1)} = 1$ if $g = \text{MAP}(\hat{\mathbf{z}}_i^{(m+1)})$ and $\tilde{z}_{i,g}^{(m+1)} = 0$ otherwise, for which the denominator in Equation (5) need not be evaluated.

Thus, a CC model fitted by CEM, with $\lambda = 1$, equal mixing proportions, and the central sequences estimated by the medoid rather than the modal sequence, is equivalent to *k*-medoids based on the Hamming distance. We leverage these similarities by applying *k*-medoids to the Hamming distance matrix in order to initialise the ECM algorithm with 'hard' starting values for the allocation matrix $\mathbf{Z}$. In particular, we rely on a weighted version of PAM available in the R package `WeightedCluster` (Studer, 2013), itself initialised using Ward's hierarchical clustering. The more closely related *k*-modes algorithm (Huang, 1997) is not used, as case-weighted implementations are currently unavailable. In any case, our strategy is less computationally onerous than using multiple random starts. Moreover, our experience suggests that the ECM algorithm converges quickly when our initialisation strategy is adopted and that a great many number of random starts are required in order to achieve comparable performance. For models with an explicit noise component, an initial guess of the prior probability τ_0 that observations are noise is required. Allocations are then initialised, assuming the last component is the one associated with $\lambda_g = 0$, by multiplying the initial $(G-1)$-column $\mathbf{Z}$ matrix by $1-\tau_0$ and appending a column in which each entry is τ_0. We caution that the initial τ_0 should not be too large.

4.3 | Model selection and validation

In contrast to heuristic clustering approaches like *k*-medoids and Ward's hierarchical method, the model-based paradigm facilitates principled model-selection using likelihood-based information criteria. In the MEDseq setting, the notion of model selection refers to identifying the optimal number of components *G* in the mixture and finding the best MEDseq model type in terms of constraints on the precision parameters. Variable selection on the subset of covariates included in the gating network can also improve the fit. For a given set of covariates, one would typically evaluate all model types over a range of *G* values and choose simultaneously both the model type and *G* value according to some criterion. Thereafter, different fits with different covariates can be compared according to the same criterion.

The Bayesian information criterion (BIC; Schwarz, 1978) includes a penalty term which depends on the number of free parameters *k* in the model. The parameter counts can be deceptive for MEDseq models. In particular, regarding the estimation of $\widehat{\theta}_{g,t}$, we note that identifying the modal state for a given time point implicitly involves estimating occurrence probabilities for $(v_t - 1)$ states and then selecting the most common. This is accounted for in Appendix A, wherein the number of free parameters under each MEDseq model type is summarised. We also note that the penalty $k \log n$ is applied to the maximum *pseudo* log-likelihood estimate in the sample-weighted setting (Xu et al., 2013).

Beyond its use in identifying the optimal *G* and precision parameter settings, the BIC is also employed in greedy stepwise selection algorithms in order to guide the inclusion/exclusion of relevant gating covariates. We propose a bidirectional search strategy in which each step can potentially consist of adding or removing a non-noise component or adding or removing a covariate. Interaction terms are not considered. Every potential action is evaluated over all possible model types at each step, rather than considering changing the model type as an action in itself. Changing the gating covariates or changing the number of components can affect the model type, as observed by Murphy and Murphy (2020). While this makes the stepwise search more computationally intensive, it is less likely to miss optimal models as it explores the model space. For steps involving both gating covariates and a noise component, models with both the GN and NGN settings can be evaluated and potentially selected.

A backward stepwise search starts from the model, with all covariates included, considered optimal in terms of the number of components *G* and the MEDseq model type. On the other hand, a forward stepwise

search uses the optimal model with no covariates included as its starting point. In both cases, the algorithm accepts the action yielding the highest increase in the BIC at each step. The computational benefits of up-weighting unique cases and discarding redundant cases are stronger for the forward search, as early steps with fewer covariates are likely to have fewer unique cases across sequence patterns and covariates.

As a means of validating the model chosen by BIC, we turn to silhouette analysis to assess the quality of the clustering in terms of internal cohesion, where high cohesion indicates high between-cluster distances and strong within-cluster homogeneity. Typically, the silhouette width is defined for clustering methods which produce a 'hard' partition (Rousseeuw, 1987), and the average silhouette width (ASW) or weighted average silhouette width (wASW; Studer, 2013) is used as a model selection criterion. However, Menardi (2011) introduces the density-based silhouette (DBS) for model-based clustering methods. This allows the 'soft' assignment information to be used, which is discarded when using the MAP assignments in the computation of the wASW. The empirical DBS for observation i is given by

$$\widehat{dbs}_i = \frac{\log\left(\frac{\hat{z}_i^0}{\hat{z}_i^1}\right)}{\max_{h\in\{1,\ldots,n\}}\left(\left|\log\left(\frac{\hat{z}_h^0}{\hat{z}_h^1}\right)\right|\right)}. \tag{6}$$

As observations are assigned to clusters via the MAP classification, $\widehat{dbs}_i$ is proportional to the log-ratio of the posterior probability associated with the MAP assignment of observation i (denoted by $\hat{z}_i^0$) to the maximum posterior probability that the observation belongs to another cluster (denoted by $\hat{z}_i^1$). Use of the MAP classification implies $0 \leq \widehat{dbs}_i \leq 1 \forall i$, with high values indicating a well-clustered data point. Ultimately, the mean or the median $\widehat{dbs}$ value can be used as a global quality measure, albeit with two modifications. First, we identify a set of crisply assigned observations having $\hat{z}_i^1$ lower than a tolerance parameter ε, here set equal to 10^{-100}. These observations are given $\widehat{dbs}_i$ values of 1 and are excluded from the computation of the maximum in the denominator of Equation (6) for reasons of numerical stability. Second, we account for the sampling weights by computing a *weighted* mean density-based silhouette criterion (wDBS). While neither the wDBS nor wASW are defined for $G = 1$, unlike the BIC, they are not employed here as model selection criteria. These silhouette summary measures are used only to validate MEDseq clustering solutions and to facilitate comparisons with other methods in Section 5.2. Higher values are preferred for both criteria.

5 | ANALYSING THE MVAD DATA

Results of fitting MEDseq models to the weighted MVAD data are provided in Section 5.1. All results were obtained via our purpose-built R package `MEDseq` (Murphy et al., 2020). The impact of discarding the sampling weights is also studied. A comparison against other approaches, including hierarchical, partitional and model-based clustering methods, is included in Section 5.2. A discussion of the insights gleaned from the solution obtained by the optimal MEDseq model is deferred to Section 6.

5.1 | Application of MEDseq

Weighted MEDseq models are fit for $G = 1, \ldots, 25$, across all 8 model types (where allowable), first with all covariates included in the gating network (again, where allowable). The noise components,

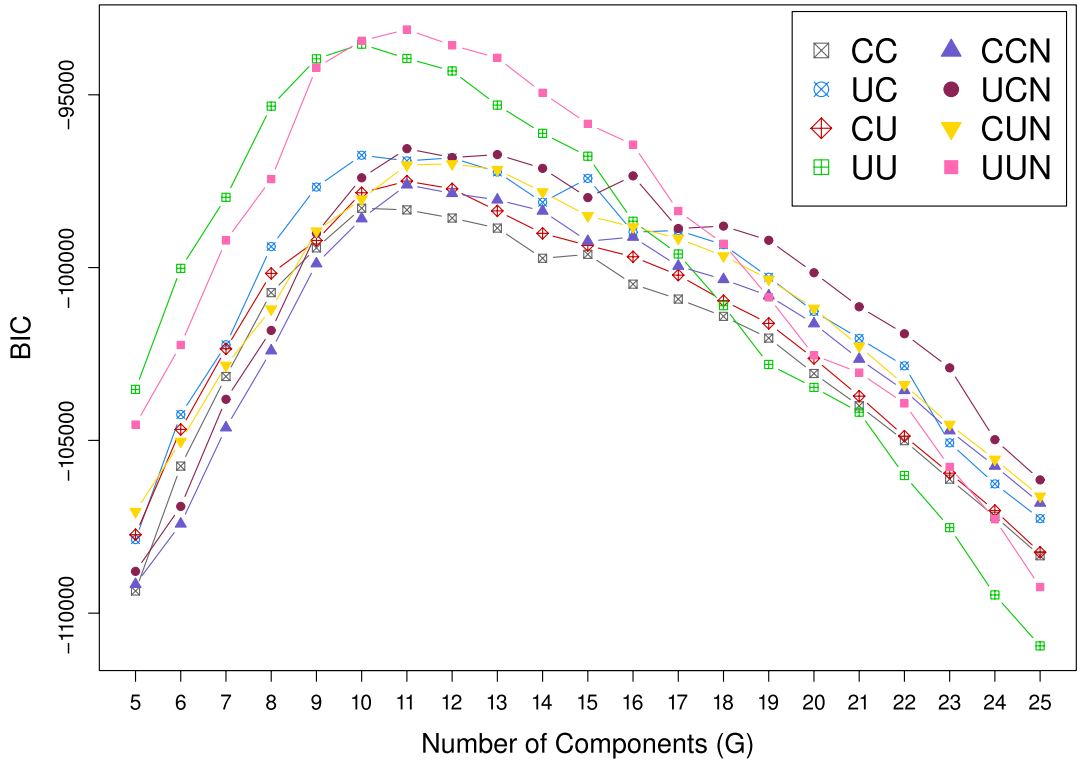

FIGURE 3 Bayesian information criterion values for all MEDseq model types, with weights and all covariates, for a range of G values [Colour figure can be viewed at wileyonlinelibrary.com]

where applicable, are treated using the NGN setting. Figure 3 shows the behaviour of the BIC for these models. To better highlight the differences in BIC, lower values for $G < 5$ are not shown. Under these conditions, a $G = 11$ UUN model is identified as optimal. The same model type and number of components are identified as optimal when the noise components are treated with the GN setting, and when the same analysis is repeated with no covariates at all.

In refining the model further via greedy stepwise selection, both the forward search (see Table 2) and backward search (see Table 3) thus begin with the same number of components and the same model type. As previously stated, covariates used to define the sampling weights are excluded in both cases. Notably, no step in either search elects to modify G or the model type. Both searches converge to the same $G = 11$ UUN model with only the single covariate 'GCSE5eq' in the NGN gating network, though the search in the forward direction does so in fewer steps. Under this model, the probability of belonging to the noise component is constant and does not depend on the included covariate.

Notably, there is little difference between the respective clusterings produced by the various models including no covariates, all covariates, and only GCSE5eq. Indeed, both the soft $\widehat{\mathbf{Z}}$ matrices and hard MAP assignments are almost identical between each pair of models; relative to the optimal model after stepwise selection, there are only 1 and 2 cases assigned to different clusters under equivalent models with no covariates and all covariates, respectively. Thus, the sequences themselves overwhelm the covariates and there is little confounding between the simultaneous roles of GCSE5eq under the optimal model in guiding both the construction of the clusters and their interpretation. Moreover, the parsimony afforded by discarding the other covariates simplifies the interpretation greatly. Thus, while adapting the 'two-step' approach introduced for LCR (Bakk & Kuha, 2018) to the MEDseq

TABLE 2 Summary of the steps taken to improve the Bayesian information criterion (BIC) in the forward direction

Optimal step	G	Model type	Gating covariates	Gating type	BIC
—	11	UUN		—	−93190.08
Add 'GCSE5eq'	11	UUN	GCSE5eq	NGN	−92953.85
Stop	11	UUN	GCSE5eq	NGN	−92953.85

TABLE 3 Summary of the steps taken to improve the Bayesian information criterion (BIC) in the backward direction

Optimal step	G	Model type	Gating covariates	Gating type	BIC
—	11	UUN	Catholic, FMPR, Funemp, GCSE5eq, Gender, Livboth	NGN	−93111.30
Remove 'FMPR'	11	UUN	Catholic, Funemp, GCSE5eq, Gender, Livboth	NGN	−93068.09
Remove 'Livboth'	11	UUN	Catholic, Funemp, GCSE5eq, Gender	NGN	−93025.73
Remove 'Catholic'	11	UUN	Funemp, GCSE5eq, Gender	NGN	−92994.32
Remove 'Funemp'	11	UUN	GCSE5eq, Gender	NGN	−92967.23
Remove 'Gender'	11	UUN	GCSE5eq	NGN	−92953.85
Stop	11	UUN	GCSE5eq	NGN	−92953.85

setting may be of interest for other applications, the results for the MVAD data do not differ greatly from those presented in Section 6, as shown in Appendix C.

For completeness, the analysis above is repeated with the sampling weights discarded entirely and consideration given where appropriate to the two covariates used to define $\mathbf{w}$. In doing so, identical inference is obtained on the model type; however, the results differ in terms of the optimal G (now 10), the uncovered partition, and the estimated model parameters. This is not surprising, as failure to account for $\mathbf{w}$ in the clustering produces biased estimates of the component-specific parameters and the cluster membership probabilities, as well as the gating network coefficients. Additionally, an extra gating covariate (Grammar) is included after stepwise selection in the unweighted analysis. However, the results are reasonably robust to a coarsening of the sequences; in repeating all analyses with the data subsetted into six-monthly intervals, similar inferences are again obtained. Notably, the ECM algorithm's runtime is not greatly reduced in doing so. Indeed, MEDseq models scale more poorly with n (or, more specifically, the number of unique cases) rather than T or v, as the number of (pseudo) likelihood evaluations required for large n is more computationally expensive than the number of simple matching evaluations required for long sequences.

5.2 | Other clustering methods

To contrast the MEDseq results for the MVAD data with those obtained by other methods, we present a non-exhaustive comparison against some distance-based and some Markovian approaches. Regarding the former, we present only some common heuristic methods which treat the distance

matrix as the input using distance metrics which are commonly adopted in the literature on life-course sequences, namely PAM and Ward's method based on the Hamming distance and OM. We note that fuzzy clustering offers an alternative distance-based perspective which also allows for soft assignments (see D'Urso (2016) for an excellent overview), with further, separate extensions for incorporating covariates and including a noise cluster in Studer (2018) and D'Urso and Massari (2013), respectively. However, this paradigm is not considered further, both for the sake of brevity and because case-weighted implementations are currently unavailable. LCA and LCR, fit via the R package `poLCA` (Linzer & Lewis, 2011), are also excluded, as they encounter computational difficulties due to the explosion in the number of parameters for $G \geq 3$. Among the considered methods, only MEDseq and the distance-based methods can accommodate the sampling weights.

First, MEDseq models with no covariates and all covariates are compared against weighted versions of *k*-medoids, using the R package `WeightedCluster` (Studer, 2013), and Ward's hierarchical clustering. Here, *k*-medoids is itself initialised using Ward's method. Neither method can be compared to MEDseq models in terms of BIC or wDBS values, as they are not model-based and do not yield 'soft' cluster membership probabilities, respectively. Thus, Figure 4 shows a comparison of wASW values using MAP classifications where necessary. Only the MEDseq model type (and gating network setting, for models with covariates) with the highest wASW for each *G* value is shown, for clarity. Note that the wASW is computed using the observed Hamming distance matrix, which both comparators in Figure 4 utilise directly, while MEDseq models are only *based on* the Hamming metric. Nonetheless, MEDseq models show superior or competitive performance across the majority of *G*

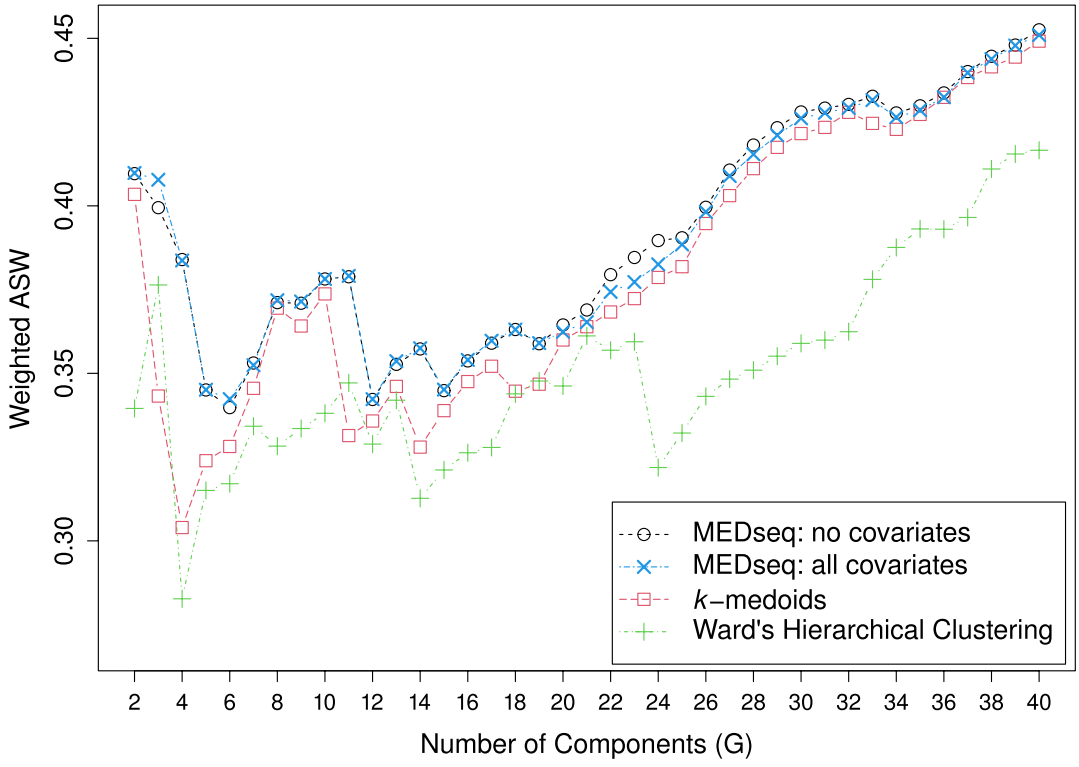

FIGURE 4 Values of the wASW measure, using Hamming distances, for the best MEDseq model type for each *G* value with no covariates and all covariates. Corresponding values for weighted versions of *k*-medoids and Ward's hierarchical clustering based on the Hamming distance are also shown [Colour figure can be viewed at wileyonlinelibrary.com]

values. In particular, the optimal model identified after stepwise selection achieves wASW = 0.386. The superior wASW values achieved by MEDseq models provide evidence that the proposed methodology, which embeds features of the distance-based approaches into a model-based setting, yields more compact and well-separated clusters. Notably, similar conclusions are drawn when OM—with the same cost settings as used in McVicar and Anyadike-Danes (2002)—is used in place of the Hamming distance for *k*-medoids and Ward's method.

Second, finite mixtures with first-order Markov components, fit via the R package `ClickClust` (Melnykov, 2016b), are also included in the comparison. This package allows the initial state probabilities to be either estimated or equal to $1/v$ for all categories; both scenarios are considered and other function arguments are set to their default values. The wASW values for the `ClickClust` models are not shown in Figure 4; they are much lower than those of the other methods up to $G = 5$ and turn negative thereafter. Though this implies inferior clustering behaviour for `ClickClust` models, the method also returns a $\hat{\mathbf{Z}}$ matrix of cluster membership probabilities. Hence, these models are also compared to MEDseq in terms of the wDBS measure in Figure 5. Again, only the best model of each type is shown for each G value; here, the MEDseq models again exhibit the best performance over the entire range. Notably, the optimal $G = 11$ UUN MEDseq model with 'GCSE5eq' in the gating network achieves wDBS = 0.455. An advantage of `ClickClust` is that it allows sequences of unequal lengths, but this is not a concern for the MVAD data.

Third, the R package `seqHMM` (Helske & Helske, 2019) provides tools for fitting mixtures of hidden Markov models, with gating covariates influencing cluster membership probabilities. Such models allow cluster memberships to evolve over time, similar to mixed membership models (Airoldi et al., 2014). They

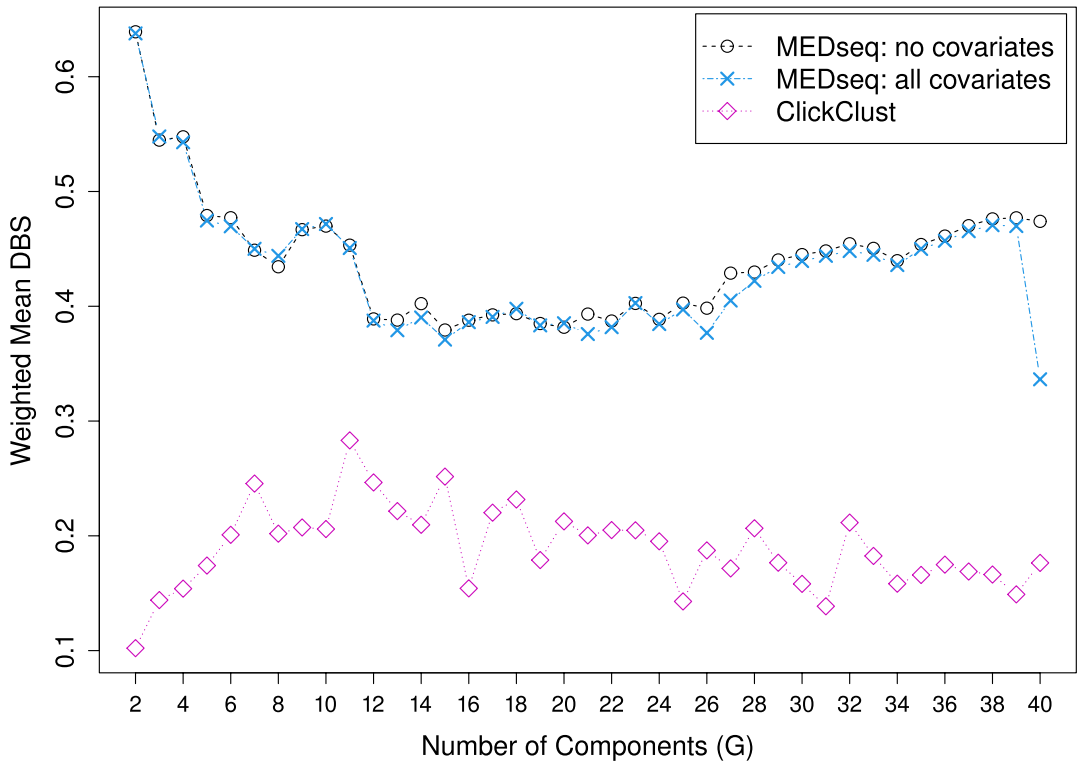

FIGURE 5 Values of the wDBS measure for the best MEDseq model type at each G value with no covariates and all covariates. Corresponding values for the best `ClickClust` model are also shown [Colour figure can be viewed at wileyonlinelibrary.com]

thus cannot be directly compared to MEDseq models. However, we note that the `seqHMM` package provides a pre-fitted model for the MVAD data, with the first 2 months also discarded and no covariates. The model has 2 clusters, with 3 and 4 hidden states, respectively, and achieves wDBS = 0.50 and wASW = 0.23. Otherwise identical `seqHMM` models, including either all covariates or only the GCSE5eq covariate chosen for the optimal MEDseq model via stepwise selection, both achieve wDBS = 0.47 and wASW = 0.23. Notably, these wDBS and wASW values are much worse than those for MEDseq models with $G = 2$. Overall, the `ClickClust` and `seqHMM` results suggest that holistic approaches—MEDseq models, in particular—yield better clusterings than Markovian ones for the MVAD data.

6 | DISCUSSION OF THE MVAD RESULTS

To better inform a discussion of the results obtained by the optimal $G = 11$ UUN model for the MVAD data, with the covariate GCSE5eq in the NGN gating network, its estimated central sequences are first shown in Figure 6. Seriation has been applied, using the observed Hamming distance matrix and the travelling salesperson combinatorial optimisation algorithm (Hahsler et al., 2008), in order to give consecutive numbers to clusters with similar estimated (weighted) modal sequences. Each cluster's label is derived from the representation of $\widehat{\boldsymbol{\theta}}_g$ in state-permanence-sequence format (SPS; Aassve et al., 2007). The same ordering and labels are used in all subsequent graphical and tabular displays of results. The uncovered clusters are shown in Figure 7, to which additional seriation has been applied in order to also group the observations within clusters, for visual clarity. Finally, the average time

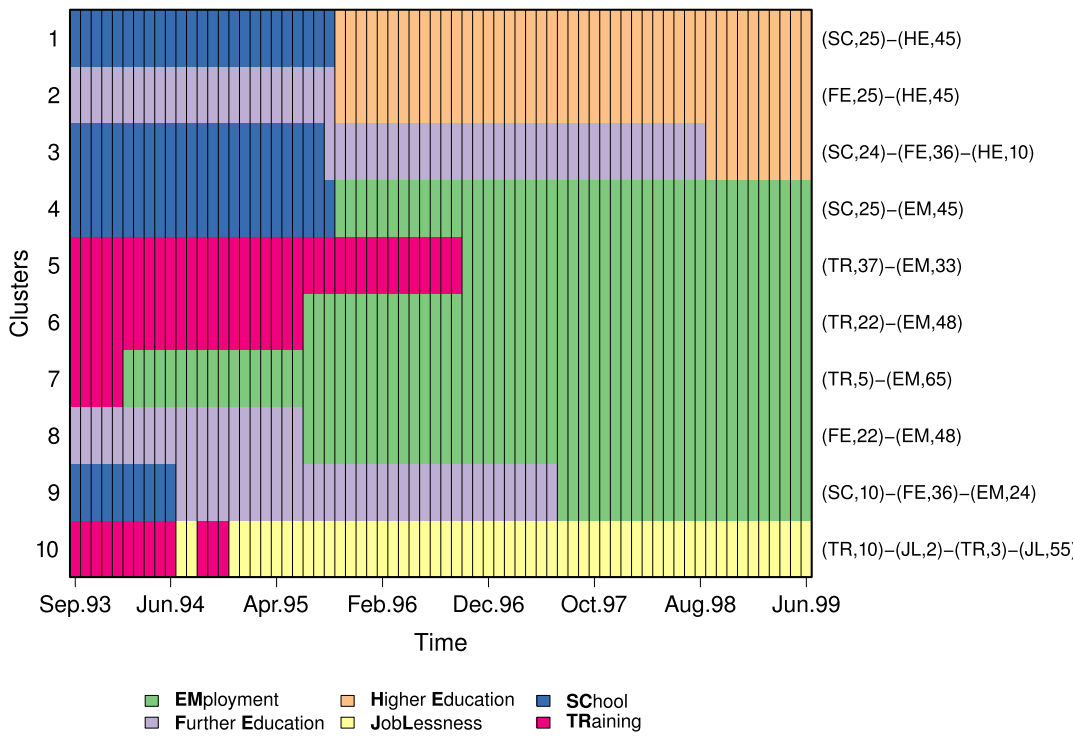

FIGURE 6 Central sequences of the optimal $G = 11$ UUN model with the GCSE5eq gating covariate. The state-permanence-sequence labels on the right characterise each non-noise cluster by the distinct successive states in $\widehat{\boldsymbol{\theta}}_g$, with associated durations (in months) [Colour figure can be viewed at wileyonlinelibrary.com]

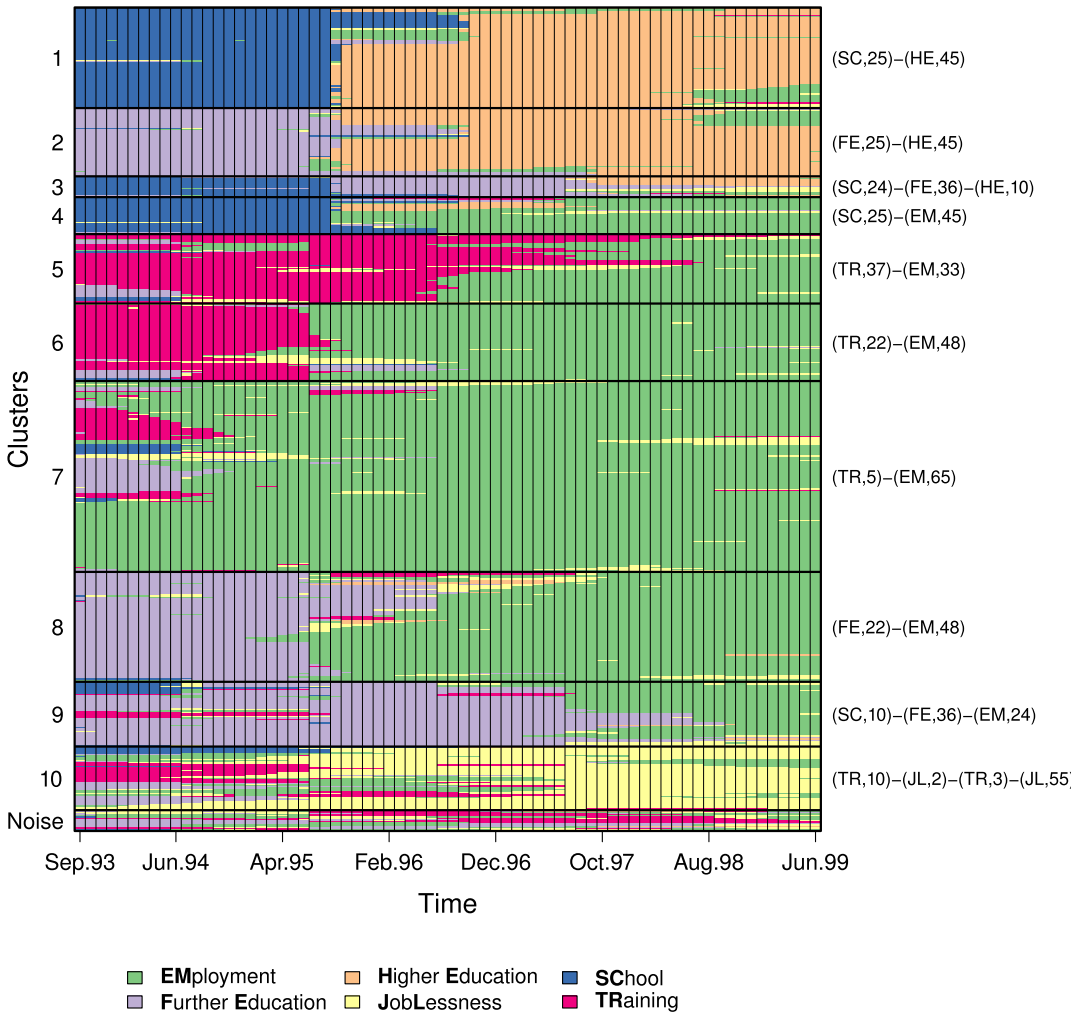

FIGURE 7 Clusters uncovered under the optimal $G = 11$ UUN model with the GCSE5eq gating covariate. The rows correspond to the $n = 712$ observed sequences, including duplicate cases previously discarded during model fitting, grouped according to the MAP classification and ordered according to the observed Hamming distance matrix. Each cluster is named according to the state-permanence-sequence representation of $\hat{\theta}_g$ [Colour figure can be viewed at wileyonlinelibrary.com]

spent in each state by cluster—weighted by w_i and the estimated cluster membership probabilities—is shown in Table 4, along with the cluster sizes.

This solution tends to group individuals who experience trajectories that are similar or that differ only for relatively short periods. In particular, the dominating combinations of states experienced over time are clearly identified, and differences in durations and/or age at transition are quite limited in size. Within clusters, substantial reduction of misalignments and/or differences in the durations of spells is evident. Ultimately, the partition is characterised not only by the sequencing (i.e. the experienced, ordered combinations of states), but also by the spell durations and the ages at transitions which appear mostly homogeneous within clusters. This can be explained by the fact that cases in the identified groups tend to dedicate the same period of time (spells of 1, 2 or 3 years) to further/higher education and/or training. This is interesting because one might expect the chosen dissimilarity metric, as it based on the Hamming distance, to attach higher importance to the sequencing.

TABLE 4 Average durations (in months) spent in each state by cluster, weighted by $\hat{z}_{i,g} w_i$, for the optimal 11-component UUN model, with the state-permanence-sequence (SPS) labels derived from $\hat{\theta}_g$. Estimated cluster sizes $\hat{n}_g$ correspond to the MAP partition

Cluster: g (SPS)		$\hat{n}_g$	EMployment	Further Education	Higher Education	Job Lessness	SChool	TRaining
1	(SC,25)-(HE,45)	87	3.77	0.29	38.45	0.89	26.07	0.54
2	(FE,25)-(HE,45)	59	4.65	26.51	37.63	0.45	0.76	0.00
3	(SC,24)-(FE,36)-(HE,10)	18	3.40	30.58	8.56	4.07	21.84	1.56
4	(SC,25)-(EM,45)	32	35.60	1.68	3.63	2.85	25.60	0.63
5	(TR,37)-(EM,33)	60	28.29	1.24	0.00	3.38	1.35	35.74
6	(TR,22)-(EM,48)	67	45.84	1.47	0.00	3.15	1.51	18.03
7	(TR,5)-(EM,65)	165	57.50	2.11	0.00	5.16	1.62	3.62
8	(FE,22)-(EM,48)	95	41.30	22.65	0.99	3.04	1.30	0.72
9	(SC,10)-(FE,36)-(EM,24)	56	21.82	35.19	0.27	3.99	6.15	2.58
10	(TR,10)-(JL,2)-(TR,3)-(JL,55)	55	8.40	3.38	0.22	42.89	4.20	10.91
Noise	—	18	21.50	11.42	1.19	14.51	2.20	19.18

The 11-cluster solution for the MVAD data separates individuals who continued in school (clusters 1, 3 and 4), or otherwise prolonged their studies after the end of compulsory education (clusters 2, 8 and 9), from those who entered the labour market (clusters 5, 6 and 7). The clear division visible for some clusters in Figure 7 around Autumn 1995, when new semesters of further and higher education commenced and the majority of those still remaining in school had eventually left, corresponds to the time point in Figure 2 after which the entropies declined. Interestingly, individuals who experienced prolonged periods of unemployment are mostly isolated in cluster 10; this is particularly important because the Status Zero Survey aimed to identify such 'at risk' subjects. From this, we conclude that youth unemployment in Northern Ireland in this period was predominantly a problem of small numbers experiencing long spells of non-participation in the labour market rather than large numbers dipping into brief, frictional spells.

Clusters 1, 3 and 4 include subjects who continued school for about 2 years, presumably to retake previously failed examinations or to pursue academic or vocational qualifications. These individuals are split into two groups depending on whether they continued their studies (FE: cluster 3, or HE: cluster 1) or were employed directly (cluster 4). Clusters 2, 8 and 9 group subjects who initially entered further education, for about 2 years (clusters 2 and 8) or more (cluster 9). Most subjects in clusters 8 and 9 entered employment directly after further education, whereas the vast majority of those in cluster 2 transitioned to higher education, where they remained until the end of the observation period.

As for the clusters of individuals who moved quickly to the labour market after the end of compulsory education, it is possible to distinguish between individuals who almost immediately found a job and remained in employment for most of the observation period (the large cluster 7) and individuals who entered government-supported training schemes (clusters 5 and 6). A further separation is between subjects who were employed after about 2 years of training (cluster 6) and those who participated in training for a much longer period (cluster 5). Importantly, most of the individuals in these two clusters were able to find a job even if some respondents experienced some periods of unemployment.

It is interesting to observe that the cluster of careers dominated by persistent unemployment (cluster 10) is characterised by different experiences at the end of the compulsory education period. Indeed, some subjects entered employment directly after the end of compulsory education but left or lost their job after some months, while some prolonged their education before becoming unemployed. However, the majority entered a training period that did not evolve into steady employment.

Notably, the optimal model identified is a UUN model, i.e. one whose precision parameters vary across both clusters and time points. Thus, model selection favours a flexible, heavily-parameterised MEDseq variant which, while based on the simple Hamming distance, has cluster-specific and period-specific costs which allow element-wise mismatches between sequences and the central sequences in different time periods in different clusters to contribute differently to the overall distance measure. While a display of the estimated precision parameters is omitted, for brevity, their values can be easily examined via the `MEDseq` R package. Nonetheless, it is already clear that the model captures different degrees of heterogeneity in the cluster-specific state distributions of each month.

The coefficients of the gating network with associated WLBS standard errors are given in Table 5, from which a number of interesting effects can be identified. The interpretation of the effects of the covariates is made clearer by virtue of there being just one included after stepwise selection. For completeness, gating network coefficients and associated WLBS standard errors for the model with all covariates included are provided in Appendix C.

Relative to the reference cluster (cluster 1), characterised by those who prolonged their schooling for two years to sit A-level exams before successfully transitioning to higher education, all slope coefficients are notably negative. All students achieving 5 or more grades at A–C in GCSE exams are therefore less likely to belong to all other clusters, relative to cluster 1. Thus, the reference level for the effect of GCSE5eq is appropriate and the interpretation is guided only by the magnitude of the slope

TABLE 5 Multinomial logistic regression coefficients and associated weighted likelihood bootstrap (WLBS) standard errors (in parentheses), with state-permanence-sequence (SPS) labels, for the non-gated noise gating network of the optimal $G = 11$ UUN model with the GCSE5eq covariate. Recall that GCSE5eq = 1 for subjects who achieved 5 or more grades at A–C (or equivalent) in GCSE exams

Cluster: *g* (SPS)		(Intercept)		GCSE5eq	
1	(SC,25)-(HE,45)	—		—	
2	(FE,25)-(HE,45)	−0.95	(0.44)	−0.47	(0.49)
3	(SC,24)-(FE,36)-(HE,10)	−0.46	(0.63)	−1.23	(0.73)
4	(SC,25)-(EM,45)	0.58	(0.44)	−2.18	(0.58)
5	(TR,37)-(EM,33)	1.03	(0.38)	−3.43	(0.55)
6	(TR,22)-(EM,48)	1.19	(0.35)	−3.73	(0.50)
7	(TR,5)-(EM,65)	1.70	(0.32)	−4.09	(0.47)
8	(FE,22)-(EM,48)	0.60	(0.38)	−2.20	(0.42)
9	(SC,10)-(FE,36)-(EM,24)	0.95	(0.39)	−3.20	(0.55)
10	(TR,10)-(JL,2)-(TR,3)-(JL,55)	0.90	(0.36)	−3.73	(0.72)

coefficients and their associated standard errors, as well as the intercepts. First, the effects for clusters 2 and 3, capturing other subjects who were in higher education by the end of the observation period, appear slight (on the basis of the size of the standard errors of their slopes). Coupled with the negative intercepts for these clusters, this suggests, as expected, that more academically inclined students tend to prolong their education in order to improve their job prospects.

Conversely, all other intercepts are positive and all other slope coefficients appear to be significantly different from 0. We can say, therefore, despite the 2-year continuation in school of subjects in cluster 4, that students who do well in GCSE exams are less likely to belong to this cluster. Furthermore, we can see the coefficient magnitudes increasing and the standard errors decreasing as we move from cluster 5 to cluster 7. As these clusters are distinguished only by the length of the training period prior to securing stable employment, this again suggests that academically poor students are quick to find a job, presumably in an unskilled capacity. Similar conclusions can be drawn for clusters 8 and 9, that is subjects who secured employment of some kind after some time in further education rather than third-level education. Finally, those who achieved 5 or more high GCSE grades are less likely to experience persistent spells of joblessness (cluster 10).

The optimal $G = 11$ UUN model contains a uniform noise component. The BIC chooses such a model over $G = 10$ models without a noise component and $G = 11$ models with all non-noise components. Detecting outliers in this way allows the remaining non-noise clusters to be modelled more clearly. Figure 8 focuses on the noise component, which isolates errant, directionless subjects who don't neatly fit into any of the defined clusters and transition quite frequently between states. This includes transitions in and out of further education, employment, and training. Most subjects here are early school-leavers. Under the model's NGN gating network, the probability of belonging to this noise component is constant (≈ 0.025) and independent of the included GCSE5eq covariate.

7 | CONCLUSION

The Status Zero Survey followed a sample of Northern Irish youths over a 6-year period, recording their employment activities at monthly intervals, in order to explore their unfolding career trajectories and

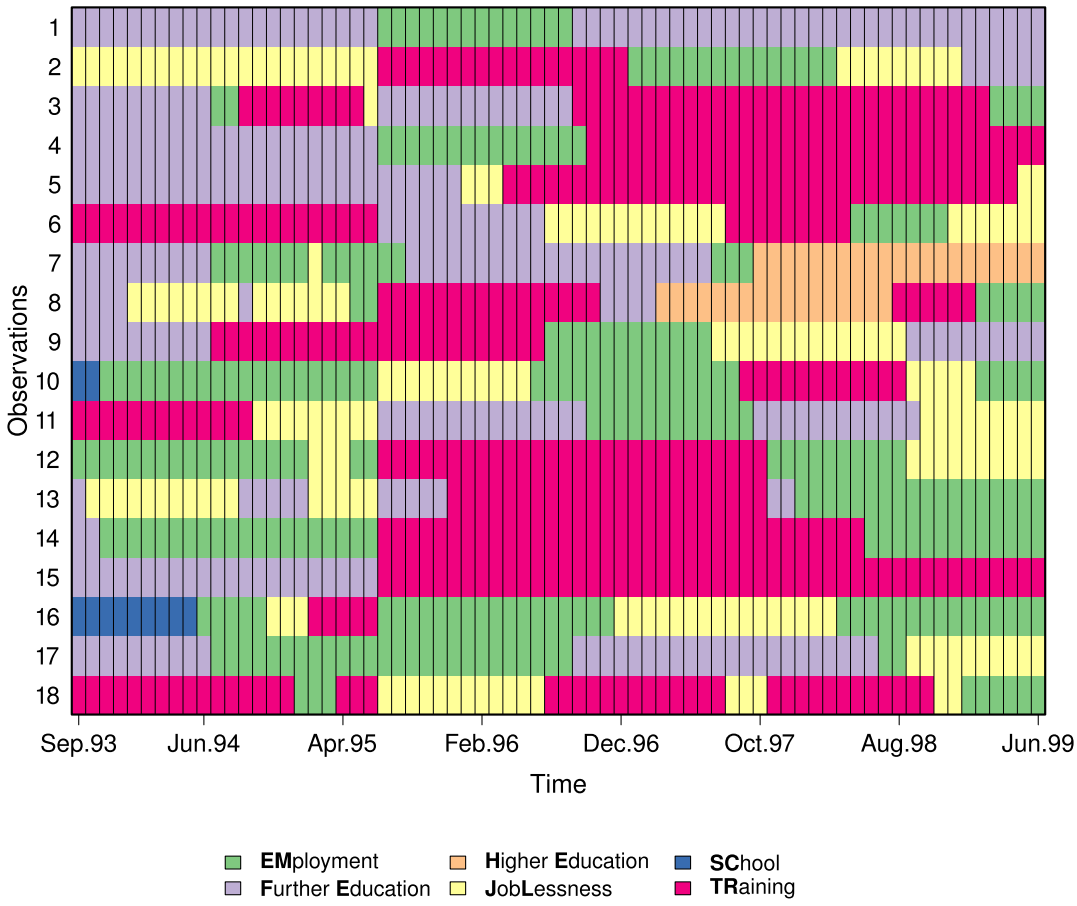

FIGURE 8 Observations assigned to the noise component of the optimal $G = 11$ UUN model with the GCSE5eq covariate in the non-gated noise gating network [Colour figure can be viewed at wileyonlinelibrary.com]

identify those at risk of prolonged unemployment. Here we present a model-based clustering approach, with the aims of assessing how many typical trajectories there are, what kinds of typical trajectories there are, and what kinds of individuals are more likely to experience which kinds of trajectories. Our approach is contrasted to heuristic approaches previously employed in analyses of these data. In McVicar and Anyadike-Danes (2002), Ward's hierarchical clustering algorithm is applied to an OM dissimilarity matrix to identify relevant patterns in the data, with subjective costs. Notably, reference is not made to the associated covariates until the uncovered clustering structure is examined. In particular, MLR is used to relate the hard assignments of the sequences to clusters to a set of baseline covariates. It is also notable that the sampling weights are incorporated only in the MLR stage and not in the clustering itself. This is arguably a three-step approach, comprising the computation of pairwise string distances using OM (or some other distance metric), the hierarchical or partition-based clustering, and the (weighted) MLR.

MEDseq models, conversely, offer a more coherent 'one-step' model-based approach. The sequences are modelled directly using a finite mixture of exponential-distance models, with the Hamming distance and weighted variants thereof employed as the distance metric. A range of precision parameter settings have been explored to allow different time points contribute differently to the overall distance. Thus, varying degrees of parsimony are accommodated. Sampling weights are accounted for by weighting each observation's contribution to the pseudo likelihood. Dependency on covariates is introduced by relating the cluster membership probabilities to covariates under the

mixture of experts framework. Thus, MEDseq models treat the weights, the relation of covariates to clusters, and the clustering itself simultaneously. Hence, MEDseq provides a coherent framework for estimating the number of clusters, identifying the relevant features of these patterns, and assessing whether these patterns are somehow influenced or shaped by the subjects' background characteristics.

Model selection in the MEDseq setting identifies a reasonable solution for the MVAD data and shows that clustering the sequences in a holistic manner allows new insights to be gleaned from these data. In particular, 11 distinct components are found, of which 10 have interpretable typical trajectories and one is an additional noise component which captures deviant cases. Thus, supported by the use of an information criterion appropriate for this model-based analysis, a more granular view of the MVAD cohort than the 5 groups uncovered in McVicar and Anyadike-Danes (2002) is provided. Furthermore, allowing for the other covariates with which the sampling weights used here are defined, GCSE exam performance at the end of the compulsory education period is found to be the most single most important predictor of cluster membership.

Opportunities for future research are varied and plentiful. Co-clustering approaches could be used to simultaneously provide clusters of the observed sequence trajectories and the time periods (Govaert & Nadif, 2013). Such an approach could be especially useful for the UUN model type identified as optimal for the MVAD data, as it would reduce the number of within-cluster period-specific precision parameters required. Indeed, parsimony has been achieved in a similar fashion in the context of finite mixtures with Markov components (Melnykov, 2016a). Additionally, grouping trajectories across time as well would enable more efficient summaries of the durations of the spells in specific states, which tend to be long for the MVAD data. In particular, using co-clustering approaches which respect the ordering of the sequences by restricting the column-wise clusters to form contingent blocks would be particularly desirable. Indeed, failure to fully account for the temporal ordering of events, due to the invariance of the Hamming distance to permutations of the time periods, is a general limitation of our framework which future work will endeavour to address.

It may also be of interest for other applications to extend the MEDseq models to accommodate sequences of different lengths, for which the Hamming distance is not defined. These different lengths could be attributable to missing data, either by virtue of sequences not starting on the same date, shorter follow-up time for some subjects, or non-response for some time points. While the Hamming distance is only defined for equal-length strings, adapting the MEDseq models to such a setting would be greatly simplified if aligning the sequences of different lengths is straightforward. Another limitation of MEDseq models is that time-varying covariates are not accommodated in the gating network. Notably, neither of these concerns are relevant for the MVAD data.

However, our analysis of the MVAD data is limited by two aspects of the gating network portion of our framework. The first substantive limitation relates to the WLBS approach used for quantifying uncertainty in the MLR coefficients. As the sampling weights arise from stratification, the standard errors obtained in this fashion are approximate. Thus, examining alternative approaches to produce fully valid variance estimates in the MEDseq setting in the presence of complex sampling designs is an interesting future research avenue.

The second limitation relates to the stepwise procedure used to identify relevant covariates. As this strategy depends on an information criterion, namely the BIC, whose penalty term is based on a parameter count, it may be prudent to relax the assumption that gating covariates must affect all components. As the number of components chosen here ($G = 11$) is moderately large, a large number of extra parameters are associated with each extra covariate (see Appendix A). Thus, only GCSE5eq is identified as optimal, as it is significantly associated with many of the typical trajectories. However, we note, for example, that Catholics are largely underrepresented in cluster 7 and largely overrepresented in cluster 10 (characterised by persistent employment and persistent joblessness, respectively) despite the omission of the covariate indicating religious affiliation from the optimal model. Incorporating

regularisation penalties into the MLR to shrink certain gating network coefficients to zero could thus be a fruitful alternative to the present stepwise covariate-selection method.

Another potential extension is to consider MEDseq models with alternative distance metrics. The distance metric in García-Magariños and Vilar (2015), which accounts for the temporal correlation in categorical sequences, is of particular interest; so, too, is OM. In general, heuristic distance-based clustering (including fuzzy methods) can more easily accommodate more sophisticated distances, while changing the MEDseq distance metric fundamentally alters the model, which needs the normalising constant and the conditional maximisation steps for parameter estimation to be tailored to the choice of metric.

MEDseq models, by virtue of being based on the Hamming distance for computational reasons, implicitly assume substitution-cost matrices with zero along the diagonal and a single value common to all other entries. The relationship between the exponent of an EDM based on the Hamming distance and the Hamming distance itself (with a common cost, typically equal to 1) is apparent from the fact that multiplying the substitution-cost matrix by any positive scalar, as per normalised variants of the Hamming distance (Elzinga, 2007; Gabadinho et al., 2011), yields the same model, because its value is absorbed into λ. This is also the case for models employing weighted Hamming distance variants under which the precision parameters, and hence the *otherwise* common substitution costs, vary across clusters and/or time points. However, all model types in the MEDseq family cannot account for situations in which some states are more different than others—for example one where the cost associated with moving from employment to joblessness is assumed to be greater than the cost associated with moving from school to training—as they assume that substitution costs are the same between each pair of states. Such concerns are most pronounced when there is an explicit ordering to the states, for example education levels (Studer & Ritschard, 2016).

Basing MEDseq on OM, for instance, would require the subjective specification, or preferably estimation, of the $v(v-1)/2$ off-diagonal entries of symmetric substitution-cost matrices. Potentially, as per the range of precision settings used for the MVAD application, the substitution-cost matrices could also be allowed to vary across clusters and/or time points. However, the normalising constant under an EDM using OM depends both on the heterogeneous substitution costs and on $\boldsymbol{\theta}$ and is unavailable in closed form, thereby greatly complicating model fitting. Indeed, dependence on $\boldsymbol{\theta}$ renders even offline pre-computation of the normalising constant infeasible for even moderately large T or v. Truncation of the sum over all sequences or importance sampling techniques could be used to address the intractability. Though not a concern for the MVAD data, as one substitution is equivalent to a deletion and an insertion for equal-length sequences, considering insertions and deletions also would present further challenges. In any case, some level of approximation would be required, while the ECM algorithm for MEDseq models based on simple matching is exact.

As well as removing the normalising constant's dependence on $\boldsymbol{\theta}$, another positive consequence of the homogeneity of substitution costs with respect to pairs of states under the Hamming distance is that the ECM algorithm used for parameter estimation scales well with the sequence length T and the size of the alphabet v, especially since such normalising constants need to be computed once, G times, or $G-1$ times *per iteration*, depending on the precision parameter settings. Though potentially restrictive, having only one parameter associated with each substitution-cost matrix, regardless of its order v, helps address concerns about overparameterisation (Studer & Ritschard, 2016), especially when the substitution costs implied by the precision parameter(s) vary across clusters and/or time points.

Furthermore, it is likely that results on the MVAD data would not differ greatly with OM used in place of the Hamming distance, particularly for models where λ varies across clusters and/or time points, save for a solution with potentially fewer clusters being found. Indeed, McVicar and Anyadike-Danes (2002) also consider a setting with *common* substitution costs and find that their results do not greatly differ from their solution with state-dependent costs. This implies that the notion that some states in the MVAD data are closer to each other than others can be questioned.

Ultimately, the UUN model adopted here preserves the timing of events, by prohibiting time-warping insertion and deletion operations, while accounting (in a cluster-specific fashion) for the timing, as well as the number, of element-wise mismatches between sequences, in such a way that all states are assumed to be equally different. Given the correspondence between Hamming distance weights, precision parameters, and implicit substitution costs in MEDseq models, it is notable that these are treated as parameters rather than inputs, and are thus estimated as part of model fitting rather than pre-specified along with the matrix of pairwise distances between sequences.

Overall, our analysis of the MVAD data provides a more granular view of the cohort of Northern Irish youths than previously available, supplemented by interpretable parameter estimates achieved through a coherent model-based framework. The MEDseq model family appears promising from the perspective of reconciling the distance-based and model-based cultures within the SA community. Indeed, the results for the MVAD data are encouraging in this respect; they seem to suggest that the unconstrained precision parameter settings adequately address the misalignment issues inherent in the use of the Hamming distance. It remains to be seen if this holds for more turbulent sequences, for example those related to employment activities tracked over longer periods.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number SFI/12/RC/2289_P2. Additionally, R. Piccarreta acknowledges the support from MIUR-PRIN 2017 project 20177BRJXS. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The authors also thank Matthias Studer and members of the Sequence Analysis Association for helpful discussions.

ORCID

Keefe Murphy http://orcid.org/0000-0002-7709-3159
T. Brendan Murphy https://orcid.org/0000-0002-5668-7046
Raffaella Piccarreta http://orcid.org/0000-0002-8876-3656
I. Claire Gormley https://orcid.org/0000-0001-7713-681X

REFERENCES

Aassve, A., Billari, F. & Piccarreta, R. (2007) Strings of adulthood: A sequence analysis of young British women's weekly work-family trajectories. *European Journal of Population*, 23(3), 369–388.

Abbott, A. & Forrest, J. (1986) Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16(3), 471–494.

Abbott, A. & Hrycak, A. (1990) Measuring resemblance in sequence data: An optimal matching analysis of musician's careers. *American Journal of Sociology*, 96(1), 145–185.

Agresti, A. (2002) *Categorical data analysis*. New York: John Wiley & Sons.

Airoldi, E.M., Blei, D.M., Erosheva, E.A. & Fienberg, S.E. (2014) *Handbook of mixed membership models and their applications*. Boca Raton: Chapman and Hall/CRC Press.

de Amorim, R.C. (2015) Feature relevance in Ward's hierarchical clustering using the L_p norm. *Journal of Classification*, 32(1), 46–62.

Armstrong, D., Istance, D., Loudon, R., McCready, S., Rees, G. & Wilson, D. (1997) 'Status 0': a socio-economic study of young people on the margin. Belfast: Training and Employment Agency, Northern Ireland Economic Research Centre.

Bakk, Z. & Kuha, J. (2018) Two-step estimation of models between latent classes and external variables. *Psychometrika*, 83(4), 871–892.

Banfield, J. & Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3), 803–821.

Billari, F.C. (2001) The analysis of early life courses: Complex description of the transition to adulthood. *Journal of Population Research*, 18(2), 119–142.

Bishop, C. M. (2006) *Pattern recognition and machine learning*. New York: Springer.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. & Lindsay, B.G. (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2), 373–388.

Bouveyron, C., Celeux, G., Murphy, T.B. & Raftery, A.E. (2019) *Model-based clustering and classification for data science: With applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Celeux, G. & Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3), 315–332.

Chambers, R.L. & Skinner, C.J. (2003) *Analysis of survey data*. Chichester: John Wiley & Sons.

D'Urso, P. (2016) Chapter 24: Fuzzy clustering. In: Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (Eds.) *Handbook of cluster analysis*. New York: Chapman and Hall, pp. 245–575.

D'Urso, P. & Massari, R. (2013) Fuzzy clustering of human activity patterns. *Fuzzy Sets and Systems*, 215, 29–54.

Dayton, C.M. & Macready, G.B. (1988) Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401), 173–178.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1), 1–38.

Elzinga, C.H. (2007) Sequence analysis: Metric representations of categorical time series. Technical Report, Department of Social Science Research Methods, Vrije Universiteit, Amsterdam.

Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011) Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.

García-Magariños, M. & Vilar, J.A. (2015) A framework for dissimilarity-based partitioning clustering of categorical time series. *Data Mining and Knowledge Discovery*, 29(2), 466–502.

Gormley, I.C. & Frühwirth-Schnatter, S. (2019) Mixtures of experts models. In: Frühwirth-Schnatter, S., Celeux, G. & Robert, C.P. (Eds.) *Handbook of mixture analysis*, Chapter 12, London: Chapman and Hall/CRC Press, pp. 279–316.

Govaert, G. & Nadif, M. (2013) *Co-Clustering: Models, algorithms and applications*. London: ISTE Wiley.

Gower, J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871.

Hahsler, M., Hornik, K. & Buchta, C. (2008) Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software*, 25(3), 1–34.

Hamming, R.W. (1950) Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2), 147–160.

Helske, S. & Helske, J. (2019) Mixture hidden Markov models for sequence data: The seqHMM package in R. *Journal of Statistical Software*, 88(3), 1–32.

Helske, S., Helske, J. & Eerola, M. (2016) Analysing complex life sequence data with hidden Markov modeling. In: Ritschard, G. & Studer, M. (Eds.), *LaCOSA II:Proceedings of international conference on sequence analysis and related methods*, pp. 209–240.

Hoos, H. & Stützle, T. (2004) *Stochastic local search: Foundations and applications*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Huang, Z. (1997) A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Lu, H., Motoda, H. & Luu, H. (Eds.) *KDD: Techniques and Applications*. Singapore: World Scientific, pp. 21–34.

Irurozki, E., Calvo, B. & Lozano, J.A. (2019) Mallows and generalized Mallows model for matchings. *Bernoulli*, 25(2), 1160–1188.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J. & Hinton, G.E. (1991) Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.

Kaufman, L. & Rousseeuw, P.J. (1990) Partitioning around medoids (program PAM). In: Kaufman, L. & Rousseeuw, P.J. (Eds.), *Finding groups in data: An introduction to cluster analysis*, Chapter 2, New York: John Wiley & Sons, pp. 68–125.

Lazarsfeld, P.F. & Henry, N.W. (1968) *Latent structure analysis*. Boston: Houghton Mifflin.

Lesnard, L. (2010) Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3), 389–419.

Levenshtein, V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.

Linzer, D.A. & Lewis, J.B. (2011) poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10), 1–29.
Mallows, C.L. (1957) Non-null ranking models. *Biometrika*, 44(1/2), 114–130.
McVicar, D. (2000) Status 0 four years on: Young people and social exclusion in Northern Ireland. *Labour Market Bulletin*, 14, 114–119.
McVicar, D. & Anyadike-Danes, M. (2002) Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(2), 317–334.
Melnykov, V. (2016a) Model-based biclustering of clickstream data. *Computational Statistics and Data Analysis*, 93(C), 31–45.
Melnykov, V. (2016b) ClickClust: An R package for model-based clustering of categorical sequences. *Journal of Statistical Software*, 74(9), 1–34.
Menardi, G. (2011) Density-based silhouette diagnostics for clustering methods. *Statistics and Computing*, 21(3), 295–308.
Meng, X.L. & Rubin, D.R. (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2), 267–278.
Muñoz-Bullón, F. & Malo, M.A. (2003) Employment status mobility from a life-cycle perspective: A sequence analysis of work-histories in the BHPS. *Demographic Research*, 9(7), 119–162.
Murphy, T.B. & Martin, D. (2003) Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis*, 41(3–4), 645–655.
Murphy, K. & Murphy, T.B. (2020) Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis and Classification*, 14(2), 293–325.
Murphy, K., Murphy, T.B., Piccarreta, R. & Gormley, I.C. (2020) `MEDseq`: mixtures of exponential-distance models with covariates. R package version 1.2.1.
O'Hagan, A., Murphy, T.B., Scrucca, L. & Gormley, I.C. (2019) Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, 34(4), 1779–1813.
Pamminger, C. & Frühwirth-Schnatter, S. (2010) Model-based clustering of categorical time series. *Bayesian Analysis*, 5(2), 345–368.
Piccarreta, R. & Studer, M. (2019) Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research*, 41, 100251.
R Core Team. (2020) R*: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20, 53–65.
Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
Studer, M. (2013) WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. Technical Report, LIVES Working Papers 24.
Studer, M. (2018) Divisive property-based and fuzzy clustering for sequence analysis. In: Ritschard, G. & Studer, M. (Eds.) *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer International Publishing, pp. 223–239.
Studer, M. & Ritschard, G. (2016) What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2), 481–511.
Ward, J. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.
Wu, L.L. (2000) Some comments on sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1), 41–64.
Xu, C., Chen, J. & Mantell, H. (2013) Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data. *Survey Methodology*, 39(2), 303–322.

How to cite this article: Murphy K, Murphy TB, Piccarreta R, Gormley IC. Clustering longitudinal life-course sequences using mixtures of exponential-distance models. *J R Stat Soc Series A*. 2021;184:1414–1451. https://doi.org/10.1111/rssa.12712

APPENDIX A

THE MEDseq MODEL FAMILY: PARAMETER COUNTS

The models in the MEDseq family differ only in their treatment of the precision parameters, which differentiate the Hamming distance and the weighted variants thereof. The BIC is used in order to choose between the 8 model types, identify the optimal G, and guide the inclusion of gating covariates. Table A1 summarises the number of free parameters k in the BIC penalty term under each MEDseq model type, in order to demonstrate the increasing level of complexity in moving from the most parsimonious CCN model to the most heavily parameterised UU model.

The number of parameters contributing to each $\widehat{\boldsymbol{\theta}}_g$ estimate notably depends on the number of states represented across all cases in each time point. Note also that parameters relating to $\widehat{\theta}_{g,t}$ corresponding to *estimated* precision parameters are counted, while those associated with fixed precision parameter values of 0 are not counted. Similarly, precision parameters estimated as 0 are counted, but precision parameters fixed at 0 associated with the noise component are not.

The number of gating network parameters is not accounted for in Table A1. When covariates are included, there are $(\boldsymbol{r}+1)\times(\boldsymbol{G}-1)$ or $(\boldsymbol{r}+1)\times(\boldsymbol{G}-2)+1$ extra parameters—under the GN and NGN settings, respectively—where $\boldsymbol{r}+1$ is the dimension of the associated design matrix, including the intercept term. When $\boldsymbol{\tau}$ is not covariate-dependent, there are $\boldsymbol{G}-1$ extra parameters when $\boldsymbol{\tau}$ is unconstrained or only 1 extra parameter if $\boldsymbol{\tau}$ is constrained and the model includes a noise component, in which case $\boldsymbol{\tau}_0$ is allowed to vary.

APPENDIX B

ESTIMATING MEDSEQ PRECISION PARAMETERS

For fixed $\boldsymbol{\theta}$, the PMF in Equation (1) belongs to the exponential family with natural parameter λ. Thus, under any distance metric, the method of moments estimate of λ is equal to the MLE. Hence, with $\widehat{\boldsymbol{\theta}}$ already estimated as per Section 4.1.2, $\widehat{\lambda}$ ensures that the expected distance of observations from $\widehat{\boldsymbol{\theta}}$ is equal to the observed average distance from $\widehat{\boldsymbol{\theta}}$, since the solution of

TABLE A1 Number of estimated parameters under each MEDseq model type. Models with names ending with the letter N, indicating the presence of a noise component for which the single precision parameter is fixed to 0, behave like the corresponding model without this component for all other components. Thus, λ and all subscript variants thereof refer here to the non-noise components only

				Number of parameters	
Model	**Precision**	λ_g **(Clusters)**	λ_t **(Time points)**	**Central sequence(s)**	**Precision**
CC	$\lambda_{g,t}=\lambda$	Constrained	Constrained	$G\sum_{t=1}^{T}(v_t-1)$	1
CCN				$(G-1)\sum_{t=1}^{T}(v_t-1)$	$\mathbb{1}(G>1)$
UC	$\lambda_{g,t}=\lambda_g$	Unconstrained	Constrained	$G\sum_{t=1}^{T}(v_t-1)$	G
UCN				$(G-1)\sum_{t=1}^{T}(v_t-1)$	$G-1$
CU	$\lambda_{g,t}=\lambda_t$	Constrained	Unconstrained	$G\sum_{t=1}^{T}(v_t-1)$	T
CUN				$(G-1)\sum_{t=1}^{T}(v_t-1)$	$\mathbb{1}(G>1)T$
UU	$\lambda_{g,t}=\lambda_{g,t}$	Unconstrained	Unconstrained	$G\sum_{t=1}^{T}(v_t-1)$	GT
UUN				$(G-1)\sum_{t=1}^{T}(v_t-1)$	$(G-1)T$

$$\frac{\partial \ell\left(\lambda|\mathbf{S},\,\widehat{\boldsymbol{\theta}},\,\mathrm{d}\right)}{n\partial\lambda}=\frac{\sum_{\boldsymbol{\sigma}\in S_v^T}\mathrm{d}(\boldsymbol{\sigma},\,\widehat{\boldsymbol{\theta}})\exp\left(-\lambda\mathrm{d}\left(\boldsymbol{\sigma},\widehat{\boldsymbol{\theta}}\right)\right)}{\sum_{\boldsymbol{\sigma}\in S_v^T}\exp\left(-\lambda\mathrm{d}\left(\boldsymbol{\sigma},\widehat{\boldsymbol{\theta}}\right)\right)}-\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}\left(\mathbf{s}_i,\,\widehat{\boldsymbol{\theta}}\right)$$

implies

$$\mathbb{E}_\lambda\left(\mathrm{d}\left(\mathbf{S},\widehat{\boldsymbol{\theta}}\right)\right)=\frac{\sum_{\boldsymbol{\sigma}\in S_v^T}\mathrm{d}\left(\boldsymbol{\sigma},\widehat{\boldsymbol{\theta}}\right)\exp\left(-\lambda\mathrm{d}\left(\boldsymbol{\sigma},\widehat{\boldsymbol{\theta}}\right)\right)}{\sum_{\boldsymbol{\sigma}\in S_v^T}\exp\left(-\lambda\mathrm{d}\left(\boldsymbol{\sigma},\widehat{\boldsymbol{\theta}}\right)\right)}=\bar{\mathrm{d}}\left(\mathbf{S},\widehat{\boldsymbol{\theta}}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}\left(\mathbf{s}_i,\,\widehat{\boldsymbol{\theta}}\right). \tag{A1}$$

Under the Hamming distance, the value of the expectation in Equation (A1) holds for any arbitrary reference sequence in place of $\widehat{\boldsymbol{\theta}}$. As the denominator in Equation (A1)—corresponding to the normalising constant in Equation (3), under the Hamming distance—is a function of λ, it is crucial that it exists in closed form in order to estimate the precision parameter. Hence, with known $\widehat{\boldsymbol{\theta}}$, the MLE for λ for an unweighted single-component CC model can be obtained as follows:

$$\ell\left(\lambda|\mathbf{S},\,\widehat{\boldsymbol{\theta}},\,\mathrm{d}_\mathrm{H}\right)=-\lambda n\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\,\widehat{\boldsymbol{\theta}}\right)-nT\log\left((v-1)\,e^{-\lambda}+1\right),$$

$$\frac{\partial\ell\,(\cdot)}{\partial\lambda}=\frac{nT\,(v-1)}{e^{\lambda}+(v-1)}-n\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\widehat{\boldsymbol{\theta}}\right),$$

$$\therefore\,\widehat{\lambda}=\log\left((v-1)\left(\frac{T}{\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\widehat{\boldsymbol{\theta}}\right)}-1\right)\right),$$

which notably relies on the inverse of the average Hamming distance normalised by the sequence length T. However, this can yield a negative value for $\widehat{\lambda}$. Recall that we only consider $\lambda\geq 0$. Since all distances are non-negative and typically not identical, $\frac{\partial\ell(\cdot)}{\partial\lambda}$ is negative $\forall\quad\lambda>0$ in the case where the sufficient statistic $\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\,\widehat{\boldsymbol{\theta}}\right)\;>\;v^{-1}T\,(v\,-\,1)$, with $\lim_{\lambda\to\infty}\frac{\partial\ell(\cdot)}{\partial\lambda}=-n\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\,\widehat{\boldsymbol{\theta}}\right)$. Thus,

$$\widehat{\lambda}=\max\left(0,\,\log\left((v-1)\left(\frac{T}{\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\widehat{\boldsymbol{\theta}}\right)}-1\right)\right)\right).$$

When $\bar{\mathrm{d}}_\mathrm{H}\left(\mathbf{S},\widehat{\boldsymbol{\theta}}\right)<v^{-1}T\,(v-1)$, such that $\widehat{\lambda}\;>\;0$, the identity $\log\,(c(a/b-1))=\log\,(c)+\log\,(a-b)-\log\,(b)$ is used for numerical stability, otherwise $\widehat{\lambda}$ is set to 0. When sampling weights are included, following the same steps as above yields the corresponding estimate

$$\widehat{\lambda}=\max\left(0,\,\log\,(v-1)\,+\log\left(\frac{Tn}{\sum_{i=1}^{n}w_i\mathrm{d}_\mathrm{H}\left(\mathbf{s}_i,\widehat{\boldsymbol{\theta}}\right)}-1\right)\right). \tag{A2}$$

The ECM algorithm is employed when $G>1$, in which case the CM-step for $\widehat{\lambda}^{(m+1)}$ under a CC MEDseq mixture model with sampling weights is given by

$$\frac{\partial \ell_c^{\mathbf{w}}(\cdot)}{\partial \lambda} = \frac{T(v-1)\sum_{i=1}^{n}\sum_{g=1}^{G} z_{i,g} w_i}{e^{\lambda}+(v-1)} - \sum_{i=1}^{n}\sum_{g=1}^{G} z_{i,g} w_i \mathrm{d}_{\mathrm{H}}\left(\mathbf{s}_i, \widehat{\boldsymbol{\theta}}_g\right),$$

$$\therefore \widehat{\lambda}^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{Tn}{\sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{z}_{i,g}^{(m+1)} w_i \mathrm{d}_{\mathrm{H}}\left(\mathbf{s}_i, \widehat{\boldsymbol{\theta}}_g^{(m+1)}\right)} - 1\right)\right). \tag{A3}$$

As per Equation (A2), this requires the current estimate of each component's central sequence. When there are no sampling weights, one need only drop the w_i terms from Equations (A2) and (A3) to estimate the precision parameters of unweighted MEDseq models. While $\widehat{\lambda}$ can potentially be estimated as zero, the inclusion of a noise component in the CCN, UCN, CUN, and UUN models makes this explicit, by restricting one cluster to have $\lambda_{g,t} = 0 \quad \forall \quad t = 1, \ldots, T$.

However, when $\widehat{\lambda}_{g,t}$ is estimated as zero rather than fixed to zero, the corresponding $\theta_{g,t}$ parameter must be estimated, as it affects the likelihood indirectly through its role in estimating the precision parameter(s). In particular—taking the UU model as an example—all state values in the t-th sequence position with non-zero $\widehat{z}_{i,g}^{(m+1)}$ are identical to $\widehat{\theta}_{g,t}^{(m+1)}$ when the corresponding denominator in Table B2 evaluates to zero, such that $\widehat{\lambda}_{g,t}^{(m+1)} \to \infty$.

Expressions for the weighted complete data pseudo likelihood functions for all model types in the MEDseq family are given in Table B1. All models are written here as though gating network covariates $\mathbf{x}_i$ are included. However, the gating networks of models with a noise component are written in the NGN form employed by the optimal model identified in Section 5.1 rather than the GN form, i.e.

TABLE B1 Weighted complete data pseudo likelihood functions for all MEDseq model types, which differ according to the constraints imposed on the precision parameters across clusters and/or time points. The expressions for the various weighted Hamming distance metric variants employed, and the associated normalising constants, are given in full

Model	Weighted complete data pseudo likelihood
CC	$\prod_{i=1}^{n}\left[\prod_{g=1}^{G}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\lambda\sum_{t=1}^{T}\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\left((v-1)e^{-\lambda}+1\right)^{T}}\right)^{z_{i,g}}\right]^{w_i}$
UC	$\prod_{i=1}^{n}\left[\prod_{g=1}^{G}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\lambda_g\sum_{t=1}^{T}\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\left((v-1)e^{-\lambda_g}+1\right)^{T}}\right)^{z_{i,g}}\right]^{w_i}$
CU	$\prod_{i=1}^{n}\left[\prod_{g=1}^{G}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\sum_{t=1}^{T}\lambda_t\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\prod_{t=1}^{T}\left((v-1)e^{-\lambda_t}+1\right)}\right)^{z_{i,g}}\right]^{w_i}$
UU	$\prod_{i=1}^{n}\left[\prod_{g=1}^{G}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\sum_{t=1}^{T}\lambda_{g,t}\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\prod_{t=1}^{T}\left((v-1)e^{-\lambda_{g,t}}+1\right)}\right)^{z_{i,g}}\right]^{w_i}$
CCN	$\prod_{i=1}^{n}\left[\left(\prod_{g=1}^{G-1}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\lambda\sum_{t=1}^{T}\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\left((v-1)e^{-\lambda}+1\right)^{T}}\right)^{z_{i,g}}\right)\left(\frac{\tau_0}{v^T}\right)^{z_{i,0}}\right]^{w_i}$
UCN	$\prod_{i=1}^{n}\left[\left(\prod_{g=1}^{G-1}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\lambda_g\sum_{t=1}^{T}\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\left((v-1)e^{-\lambda_g}+1\right)^{T}}\right)^{z_{i,g}}\right)\left(\frac{\tau_0}{v^T}\right)^{z_{i,0}}\right]^{w_i}$
CUN	$\prod_{i=1}^{n}\left[\left(\prod_{g=1}^{G-1}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\sum_{t=1}^{T}\lambda_t\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\prod_{t=1}^{T}\left((v-1)e^{-\lambda_t}+1\right)}\right)^{z_{i,g}}\right)\left(\frac{\tau_0}{v^T}\right)^{z_{i,0}}\right]^{w_i}$
UUN	$\prod_{i=1}^{n}\left[\left(\prod_{g=1}^{G-1}\left(\tau_g(\mathbf{x}_i)\frac{\exp\left(-\sum_{t=1}^{T}\lambda_{g,t}\mathbb{1}(s_{i,t}\neq\theta_{g,t})\right)}{\prod_{t=1}^{T}\left((v-1)e^{-\lambda_{g,t}}+1\right)}\right)^{z_{i,g}}\right)\left(\frac{\tau_0}{v^T}\right)^{z_{i,0}}\right]^{w_i}$

TABLE B2 CM-steps for the precision parameter(s) of the non-noise components for all MEDseq model types, which differ according to the constraints imposed across clusters and/or time points

Model	Precision parameter CM-steps
CC	$\widehat{\lambda}^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{Tn}{\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{i,g}^{(m+1)} w_i \mathrm{d_H}\left(\mathbf{s}_i, \widehat{\boldsymbol{\theta}}_g^{(m+1)}\right)} - 1\right)\right)$
UC	$\widehat{\lambda}_g^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{T\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i}{\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i \mathrm{d_H}\left(\mathbf{s}_i, \widehat{\boldsymbol{\theta}}_g^{(m+1)}\right)} - 1\right)\right)$
CU	$\widehat{\lambda}_t^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{n}{\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{i,g}^{(m+1)} w_i \mathbb{1}(s_{i,t} \neq \widehat{\theta}_{g,t}^{(m+1)})} - 1\right)\right)$
UU	$\widehat{\lambda}_{g,t}^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i}{\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i \mathbb{1}(s_{i,t} \neq \widehat{\theta}_{g,t}^{(m+1)})} - 1\right)\right)$
CCN	$\widehat{\lambda}^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{T\sum_{i=1}^{n}\sum_{g=1}^{G-1}\widehat{z}_{i,g}^{(m+1)} w_i}{\sum_{i=1}^{n}\sum_{g=1}^{G-1}\widehat{z}_{i,g}^{(m+1)} w_i \mathrm{d_H}\left(\mathbf{s}_i, \widehat{\boldsymbol{\theta}}_g^{(m+1)}\right)} - 1\right)\right)$
UCN	$\widehat{\lambda}_g^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{T\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i}{\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i \mathrm{d_H}\left(\mathbf{s}_i, \widehat{\boldsymbol{\theta}}_g^{(m+1)}\right)} - 1\right)\right)$
CUN	$\widehat{\lambda}_t^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{\sum_{i=1}^{n}\sum_{g=1}^{G-1}\widehat{z}_{i,g}^{(m+1)} w_i}{\sum_{i=1}^{n}\sum_{g=1}^{G-1}\widehat{z}_{i,g}^{(m+1)} w_i \mathbb{1}(s_{i,t} \neq \widehat{\theta}_{g,t}^{(m+1)})} - 1\right)\right)$
UUN	$\widehat{\lambda}_{g,t}^{(m+1)} = \max\left(0, \log(v-1) + \log\left(\frac{\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i}{\sum_{i=1}^{n}\widehat{z}_{i,g}^{(m+1)} w_i \mathbb{1}(s_{i,t} \neq \widehat{\theta}_{g,t}^{(m+1)})} - 1\right)\right)$

it is assumed that τ_0 is constant, meaning the covariates do not affect the probability of belonging to the noise component (see Section 3.4).

Table B2 outlines the corresponding CM-steps for the precision parameter(s). All derivations closely follow the same steps as in Equation (A3) for the CC model and the normalised sampling weights are accounted for in all cases. These formulas can be simplified somewhat for unweighted models and/or models without gating covariates. Recall that the first letter of the model name denotes whether the precision parameters are constrained/unconstrained across clusters, the second denotes the same across time points (i.e. sequence positions), and model names ending with the letter N include a noise component.

APPENDIX C

MVAD DATA: GATING NETWORK COEFFICIENTS

Multinomial logistic regression coefficients and associated WLBS standard errors of the NGN gating network of a $G = 11$ UUN model with the GCSE5eq covariate are provided in Table 5. For completeness, coefficients and standard errors for an otherwise equivalent model with all covariates included (except those used to define the sampling weights) are given in Table C1. Such a model achieves a BIC value of −93111.30 (see Table 3), compared to −92953.85 for the optimal model with only the GCSE5eq covariate detailed in Section 5.1. Notably, $G = 11$ and the UUN model type are both also optimal for the model with all covariates included. Furthermore, both models yield identical SPS labels, derived from $\widehat{\boldsymbol{\theta}}_g$. Thus, Table 5 and Table C1 share the same reference level. In the latter case, we caution that the covariates 'Livboth' and 'Funemp' were measured after the observation period had

TABLE C1 Multinomial logistic regression coefficients and associated weighted likelihood bootstrap (WLBS) standard errors (in parentheses) for the non-gated noise gating network of the $G = 11$ UUN model with all covariates included

Cluster: g (SPS)		(Intercept)	Catholic	FMPR	Funemp	GCSE5eq	Gender	Livboth
1	(SC,25)-(HE,45)	—	—	—	—	—	—	—
2	(FE,25)-(HE,45)	−0.97 (0.60)	−0.65 (0.41)	−0.49 (0.42)	0.15 (0.76)	−0.44 (0.50)	−0.28 (0.41)	0.73 (0.45)
3	(SC,24)-(FE,36)-(HE,10)	−0.70 (1.09)	−0.14 (0.79)	−0.03 (0.70)	−1.23 (1.62)	−1.42 (0.72)	−0.56 (0.68)	0.95 (0.78)
4	(SC,25)-(EM,45)	1.43 (0.70)	−0.54 (0.54)	−0.50 (0.61)	0.94 (0.95)	−2.15 (0.56)	−0.71 (0.56)	−0.45 (0.56)
5	(TR,37)-(EM,33)	0.78 (0.60)	−0.51 (0.46)	−0.98 (0.56)	0.48 (0.73)	−3.24 (0.54)	1.63 (0.53)	−0.70 (0.50)
6	(TR,22)-(EM,48)	1.27 (0.67)	−0.50 (0.46)	−0.64 (0.48)	0.39 (0.78)	−3.60 (0.54)	0.63 (0.46)	−0.12 (0.49)
7	(TR,5)-(EM,65)	2.48 (0.53)	−1.10 (0.39)	−1.39 (0.50)	0.79 (0.71)	−3.89 (0.49)	0.33 (0.38)	−0.48 (0.42)
8	(FE,22)-(EM,48)	1.24 (0.61)	−0.06 (0.41)	−0.61 (0.44)	0.46 (0.71)	−2.15 (0.46)	−0.21 (0.40)	−0.64 (0.42)
9	(SC,10)-(FE,36)-(EM,24)	1.35 (0.71)	−0.81 (0.60)	−0.06 (0.68)	1.43 (0.94)	−3.10 (0.57)	−0.11 (0.54)	−0.35 (0.52)
10	(TR,10)-(JL,2)-(TR,3)-(JL,55)	0.65 (1.54)	0.13 (0.88)	−0.47 (0.61)	1.86 (0.91)	−3.68 (0.89)	−0.15 (0.70)	−0.11 (0.70)

begun, which complicates interpretation. In particular, subjects' living arrangements were recorded after two years, and the father's employment status was recorded in the final month of June 1999.

The MLR coefficients in Table 5 are estimated under a one-step approach, under which clustering and the relation of clusters to covariates are performed simultaneously. There are subtle interpretational differences between covariates serving as predictors of cluster-membership under a one-step approach compared to covariates enabling interpretation of the type of observation characterising each cluster under a two-step approach. Fortunately, the MEDseq framework permits both types of analysis. Thus, we present coefficients obtained under two 2-step approaches in Table C2 in order to contrast them to those in Table 5.

Firstly, a model without any covariates which is otherwise exactly equivalent to the optimal one including only GCSE5eq as a covariate is fitted (BIC = −93190.08). Thereafter, MLR is used (with the GCSE5eq covariate, with the same reference level) firstly with the 'soft' $\widehat{\mathbf{Z}}$ matrix as the response and secondly with the hard MAP partition used as the response. The sampling weights are accounted for in both cases; so too is the removal of the noise component, as per the NGN gating network setting shown in Table 5, by appropriately renormalising $\widehat{\mathbf{Z}}$ prior to estimating each MLR (i.e. prior to constructing the MAP partition under the hard approach). Notably, there is little difference between the coefficients obtained under the one-step and soft two-step approaches. However, the estimates differ more greatly under the hard two-step approach, which suggests that relating covariates to hard partitions is inappropriate when the clusters are insufficiently homogeneous, as the hard partitions do not necessarily lead to appropriate responses for the MLR.

Notably, similar conclusions are drawn when the coefficients in Table C1 are compared to those obtained (but not shown here) when similar soft and hard two-step approaches are used with the same set of 'all' covariates; namely, the coefficients differ little between the one-step and soft two-step approach, but differ greatly between the one-step and hard two-step approach. However, the coefficients differ only in magnitude and not in sign in all pairwise comparisons.

TABLE C2 MLR coefficients obtained under two alternative two-step estimation approaches for the optimal model for the MVAD data. The first and second pair of rows give the intercept and slope coefficients when the soft cluster membership probabilities and hard assignments are used as the response, respectively

Cluster: *g*		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Responses: $\widehat{\mathbf{z}}_i$	(Intercept)	—	−0.95	−0.38	0.58	1.02	1.18	1.70	0.60	0.93	0.90
	GCSE5eq	—	−0.47	−1.31	−2.18	−3.43	−3.72	−4.07	−2.21	−3.18	−3.77
Responses: $\text{MAP}\left(\widehat{\mathbf{z}}_i\right)$	(Intercept)	—	−1.01	−0.46	0.58	0.96	1.18	1.69	0.61	0.91	0.87
	GCSE5eq	—	−0.45	−1.24	−2.18	−3.66	−3.71	−4.30	2.21	−3.24	−4.12