Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions

Data Mining and Knowledge Discovery (2022) 36:379–413
https://doi.org/10.1007/s10618-021-00808-x

Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions

Rianne M. Schouten[1] · Marcos L. P. Bueno[1] · Wouter Duivesteijn[1] · Mykola Pechenizkiy[1]

Received: 31 January 2021 / Accepted: 13 October 2021 / Published online: 24 November 2021

Abstract
Discrete Markov chains are frequently used to analyse transition behaviour in sequential data. Here, the transition probabilities can be estimated using varying order Markov chains, where order k specifies the length of the sequence history that is used to model these probabilities. Generally, such a model is fitted to the entire dataset, but in practice it is likely that some heterogeneity in the data exists and that some sequences would be better modelled with alternative parameter values, or with a Markov chain of a different order. We use the framework of Exceptional Model Mining (EMM) to discover these exceptionally behaving sequences. In particular, we propose an EMM model class that allows for discovering subgroups with transition behaviour of varying order. To that end, we propose three new quality measures based on information-theoretic scoring functions. Our findings from controlled experiments show that all three quality measures find exceptional transition behaviour of varying order and are reasonably sensitive. The quality measure based on Akaike's Information Criterion is most robust for the number of observations. We furthermore add to existing work by seeking for subgroups of sequences, as opposite to subgroups of transitions. Since we use sequence-level descriptive attributes, we form subgroups of entire sequences, which is practically relevant in situations where you want to identify the originators of excep-

Responsible editor: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann

✉ Rianne M. Schouten
r.m.schouten@tue.nl

Marcos L. P. Bueno
m.l.de.paula.bueno@tue.nl

Wouter Duivesteijn
w.duivesteijn@tue.nl

Mykola Pechenizkiy
m.pechenizkiy@tue.nl

[1] Eindhoven University of Technology, Eindhoven, The Netherlands

tional sequences, such as patients. We show this relevance by analysing sequences of blood glucose values of adult persons with diabetes type 2. In the experiments, we find subgroups of patients based on age and glycated haemoglobin (HbA1c), a measure known to correlate with average blood glucose values. Clinicians and domain experts confirmed the transition behaviour as estimated by the fitted Markov chain models.

Keywords Exceptional Model Mining · Markov chains · Information-theoretic scoring functions · Sequential medical data

1 Introduction

Markov models in all their variants are frequently used to mine patterns in sequential data. Consider, for instance, 1st order Markov chains (Wilks 1999; Pirolli and Pitkow 1999; Sarukkai 2000), Hidden Markov Models (HMM) (Jaroszewicz 2010; Peharz et al. 2014; Meier et al. 2015; Bueno et al. 2019) and Dynamic Bayesian Networks (DBN) (Dagum et al. 1992; Bueno et al. 2020). All these models are called *memoryless* if they satisfy the *Markov property*: given the data at time $t-1$, the data at time t is independent of the data before time $t-1$. Furthermore, a model is *homogeneous* if its parameters do not change over time, and the sequences are *stationary* if the initial values follow the same model (Zucchini et al. 2017).

We consider discrete Markov chains where the observations are discrete values, or states, from a countable set which is called the state-space. Generally, such a model is fitted to the entire dataset and the parameter estimates give information about the average transition behaviour between states. However, some heterogeneity in the data often likely exists, and hence some sequences would be better modelled separately. We use Exceptional Model Mining (EMM) (Leman et al. 2008; Duivesteijn et al. 2016) to discover these exceptionally behaving sequences.

EMM is a local pattern mining technique seeking subsets of the dataset that behave somehow exceptionally. Here, exceptional behaviour is measured in terms of parameters of a *model class* over target attributes. A *quality measure* quantifies this exceptionality (see Sect. 2). Since EMM allows for ≥ 2 attributes to be part of the target model, it can be seen as a generalisation of Subgroup Discovery (SD) (Klösgen 1996; Wrobel 1997; Herrera et al. 2011), which uses 1 target attribute. Both frameworks employ a rule-based description language where resulting subgroups are described as a conjunction of attribute-value pairs.

An EMM model class exists for 1st order Markov chains (Lemmerich et al. 2016). We extend their work by considering Markov chains of varying order, where order k specifies the length of the sequence that is used as *memory* in the model. Specifically, our method allows for discovering subgroups in situations where the order of the Markov chain differs between the subgroup and the dataset. This situation requires comparing unequal numbers of parameters. Hence, we do not use a parameter-based quality measure, as is common in EMM, but show how information-theoretic scoring functions can evaluate a subgroup's exceptionality.

We furthermore add to existing work by seeking subgroups of sequences, as opposed to subgroups of transitions (Lemmerich et al. 2016). Whereas the latter detects het-

erogeneity within sequences, we find subgroups of homogeneous sequences that are heterogeneous w.r.t. the entire dataset. Our model class is practically relevant for identifying the originator of an exceptional sequence, such as a patient with exceptional blood glucose fluctuations (see Sect. 6.1) or an atypical user session in click-stream data (f.e. Sadagopan and Li 2008).

In sum, our main contributions include (1) an EMM model class for detecting exceptional transition behaviour of varying order, (2) a new set of quality measures based on information-theoretic scoring functions and (3) an understanding of how descriptive attributes can be used to form subgroups of entire sequences. In the rest of this paper, we introduce EMM and Markov chains in Sect. 2 and discuss related work in Sect. 3. We then propose our methodology in Sect. 4. In Sect. 5 we present our findings from controlled experiments and in Sect. 6 we analyse real-world data. Finally, Sect. 7 contains a discussion and Sect. 8 concludes.

2 Background

In the following, we will explain theoretical concepts by means of the DIALECT-2 study (Gant et al. 2017). DIALECT-2 is an observational study of adult persons with diabetes type 2, where blood glucose is measured every 15 minutes for a period of 14 days. We discretise the continuous measurements into five blood glucose levels: below range 2 (BR_2), below range 1 (BR_1), in range (IR), above range 1 (AR_1), and above range 2 (AR_2), where range refers to desired blood glucose values. We then analyse transition patterns between these levels. As a running example to aid illustration of concepts introduced in the subsequent sections, some DIALECT-2 transition patterns can be found in Fig. 1; more details on the study and its interpretation are provided in Sect. 6.1.

2.1 Preliminaries

Assume a dataset Ω with N independently but not identically distributed sequences of discrete random variables $X_t : t \in \{1, 2, ..., T\}$. The data realisation at time t is denoted with x_t. Although sequence $r \in \Omega$ has length T_r, without loss of generality, we assume one fixed length T for every sequence. We refer to N as the data *size* and write M to denote the total number of *observations*, where $M = \sum_{r \in \Omega} T_r = NT$ (note that the total number of *transitions* is $M - N$). The set of possible discrete values is $V = \{v_1, v_2, ..., v_S\}$ for all x_t. For instance, in the DIALECT-2 dataset, $N = 126$, $T = 1344$, $M \approx 170000$, $V = \{BR_1, BR_2, IR, AR_1, AR_2\}$ and $S = 5$.

We assume the availability of an extra set of attributes with information about the sequences: the descriptive attributes. The full form of sequence r then becomes $(x_1, x_2, ..., x_{T_r}, a_1, a_2, ..., a_m)$ for all $r \in \Omega$. Here, m simply denotes the number of descriptive attributes. Depending on the application, these attributes could describe personal or medical characteristics such as HbA1c category, duration of the illness and BMI (Fig. 1b), user session information such as browser language and timezone (if the sequences are click-streams) or contain meta-information about the sequences

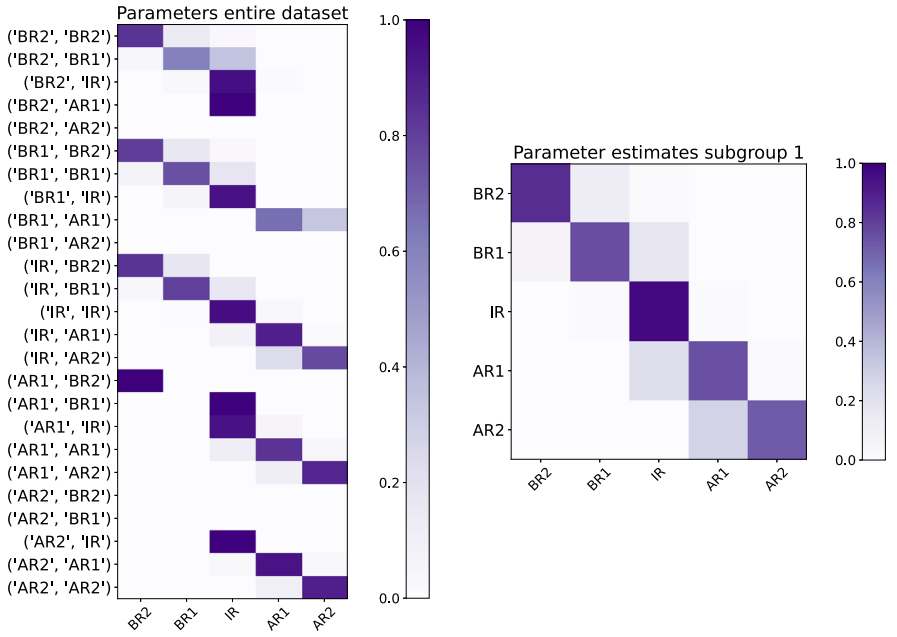

Fig. 1 Transition patterns of blood glucose levels. (a) The entire dataset follows a 2nd order Markov chain. (b) The top subgroup follows a 1st order Markov model

such as its length, state-space and starting time (see Sect. 6.2). We now explain how these descriptive attributes are used to form subgroups.

2.2 Exceptional model mining

Exceptional Model Mining (EMM) is a local pattern mining framework, seeking subgroups in a population that behave somehow exceptionally (Leman et al. 2008; Duivesteijn et al. 2016). Those subgroups have interpretable descriptions and explainable circumstances under which exceptional behaviour occurs. In short, EMM splits the data attributes into two distinct sets. The exceptionality of a candidate subgroup is gauged with a quality measure, exploring a model class over *target attributes*. We will further explain quality measures and model classes in Sect. 2.2.2. *Descriptive attributes* are used to form interpretable subgroups by deploying a rule-based description language using conjunctions of attribute-value pairs. For instance, the subgroup in Fig. 1b is described by *HbA1c category* = *low* $\wedge$ *diabetes duration* $\leq$ *20 years* $\wedge$ *21.3* $\leq$ *BMI* $\leq$ *35.7*.

2.2.1 Defining subgroups: descriptions

Denoting the collective domain of the descriptive attributes $(a_1, a_2, \ldots, a_m)$ by $\mathcal{A}$, a description is formally defined as a function $D : \mathcal{A} \rightarrow \{0, 1\}$. Subsequently, a sequence r is covered by description D if and only if $D(a_1^r, a_2^r, \ldots, a_m^r) = 1$.

Definition 1 (**Subgroup**) The subgroup corresponding to a description D is the bag of sequences $SG_D \subseteq \Omega$ that D covers:

$$SG_D = \{r \in \Omega \mid D(a_1^r, a_2^r, \ldots, a_m^r) = 1\}$$

It is important to realise that we define a subgroup such that a sequence is either covered by a description or not (we thus replaced the word *record* in Definition 1 of Duivesteijn et al. (2016) with the word *sequence*). We do not allow for a sequence to be split into pieces and to be partly assigned to a subgroup. The reason is that we want to find the originators of exceptional sequences because this could assist domain experts in adopting appropriate policies. For instance, an interpretable description of patients could assist doctors in selecting the most useful treatment; descriptions that only partly include certain patients are less helpful.

The strict partitioning between target and descriptive attributes is a powerful feature in EMM, allowing us to form subgroups of independently distributed cases while analysing sequential patterns. We are thus able to make subgroups of entire sequences because the descriptive attributes contain sequence-level information (see Sect. 2.1). In contrast, Lemmerich et al. (2016) find subgroups of transitions, using descriptive information on the transition (or time) level.

2.2.2 Evaluating exceptionality: quality measures

A quality measure quantifies the difference between behaviour within the subgroup and behaviour within the entire dataset (or the subgroup's complement). The choice of model over the target attributes is called the model class, and the quantification of the difference is given by a quality measure. The challenge in EMM is to effectively search through the descriptive space to find the top-q best-scoring subgroups.

Definition 2 (**Quality Measure**) A quality measure is a function $\varphi : \mathcal{D} \rightarrow \mathbb{R}$ that assigns a numerical value to a description D.

Generally, quality measures directly compare one or more parameter estimates, such as the difference between estimated slopes in a regression model (Duivesteijn et al. 2012) or the difference between estimated correlations of two target attributes (Duivesteijn et al. 2016). Following the terminology of Song (2017), we call these quality measures *parameter-based*. Their advantage is that you immediately know why a resulting subgroup is exceptional. However, a parameter-based approach also restricts the model in the subgroup to have the same number of parameters as the model in the entire dataset. In case of Markov chains, for instance, parameter-based quality measures would not allow the subgroup to be fitted with a higher (or lower) order model than the one that is fitted in the entire dataset. In this paper, we therefore propose to

evaluate a subgroup's exceptionality using quality measures based on information-theoretic scoring functions. We call these quality measures *evaluation-based*. For instance, in the DIALECT-2 study, the entire dataset is best modelled with a second-order Markov chain, as illustrated in Fig. 1a, while the top subgroup is best modelled with a first-order Markov chain, as illustrated in Fig. 1b. Our evaluation-based quality measures can gauge the exceptionality of the difference between these two models and their respective transition probabilities.

2.3 Markov chains

We will now first introduce 1st order Markov chains, and then extend the principles to k^{th} order chains. In this section, we overload the symbol Ω to refer only to the target attributes x_t for all $t \in \{1, 2, ..., T\}$ and temporarily forget about the descriptive attributes. We write SG to denote the same set of attributes for the subgroup.

Using the product rule and the Markov property that given the data at time $t-1$, the data at time t is independent of the data before time $t-1$, the joint probability distribution of Ω is modelled with a 1st order Markov chain by

$$P(\Omega|\theta) = P(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_T|\pi, \mathbf{A}) = p(\mathrm{x}_1)\prod_{t=2}^{T} p(\mathrm{x}_t|\mathrm{x}_{t-1}). \tag{1}$$

The prior distribution $p(\mathrm{x}_1)$ is parameterised with an *initial probabilities vector* $\pi = [\pi_1, ..., \pi_S]$. The main interest is in the transition behaviour between time $t-1$ and t, which is parameterised with an $S \times S$ probability matrix denoted with $\mathbf{A}$. Parameters $\alpha_{ij} \in \mathbf{A}\ \forall i, j \in \{1, 2, ..., S\}$ are estimated using Maximum Likelihood Estimation (MLE) where

$$p(\mathrm{x}_t = v_j|\mathrm{x}_{t-1} = v_i) = \alpha_{ij} = \frac{n_{ij}}{\sum_{j=1}^{S} n_{ij}}. \tag{2}$$

Here, n_{ij} denotes the total number of transitions from source state v_i to target state v_j $\forall i, j \in \{1, 2, ..., S\}$. Eq. (2) thus practically means that we first calculate a transition frequency matrix, and then calculate the probabilities by dividing by the sum of each row. Consequently, $\forall_i \sum_{j=1}^{S} \alpha_{ij} = 1$.

Figure 1b shows such a 1st order transition probability matrix. The dark purple square in the top left corner expresses the probability (which is $\alpha_{11} = 0.85$) that the next blood glucose level is BR_2 (column) given that the current blood glucose level is BR_2 (row). The high probabilities on the diagonal indicate that patients are likely to stay at the same blood glucose level, although patients with a current blood glucose level of AR_2 are also quite likely to transition to a lower blood glucose value (to AR_1, $\alpha_{54} = 0.28$).

The Markov chain model in Eq. (1) assumes *homogeneous* sequences where the transition parameters do not change over time (Zucchini et al. 2017). If we additionally assume that the initial probabilities vector π follows that same transition model, we say the sequences are *stationary* (Zucchini et al. 2017). It makes sense to make both

assumptions together. After all, if we assume that the transition behaviour does not change between time points 1 and T, it does not matter where or when the sequence starts. For example, if we estimate that 30% of the sequences move from state v_i to state v_j, it is also likely that 30% of the sequences start with state v_i. As a consequence, it is not necessary to separately estimate the parameters in π. Instead, we derive the initial probabilities by normalising over all j target states in the frequency matrix. The total number of free parameters in a 1st order Markov chain is therefore $K = S(S-1)$.

However, depending on the application or for very short sequences, the starting point of the sequence could be of separate interest. Consider, for example, a subgroup of patients who present themselves with different symptoms than the overall patient population. In that case, the parameters in π are separately estimated using only the data from the first time point,

$$p(\mathrm{x}_1 = v_h) = \pi_h = \frac{n_h^{(t=1)}}{\sum_{h=1}^{S} n_h^{(t=1)}}. \tag{3}$$

Here, we indicate the selection of time points in the superscript ($t = 1$). Separately estimating initial probabilities would add another $S-1$ free parameters to the Markov chain model.

Extending the 1st order Markov chain model to a kth order model gives the following joint probability distribution,

$$P(\Omega|\theta) = P(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_T|\theta) = p(\mathrm{x}_1) \cdot p(\mathrm{x}_2|\mathrm{x}_1) \cdot ... \cdot p(\mathrm{x}_k|\mathrm{x}_{k-1}, ..., \mathrm{x}_2, \mathrm{x}_1) \cdot \prod_{t=k+1}^{T} p(\mathrm{x}_t|\mathrm{x}_{t-k}, \mathrm{x}_{t-k+1}, ..., \mathrm{x}_{t-2}, \mathrm{x}_{t-1}).$$

Such a model uses the memory of time $t-1, t-2, ..., t-k$ to predict the state value at time t. To understand the transition matrix, it can be helpful to consider the k-length history as one time point with S^k possible states. Transition matrix ${}_k\mathbf{A}$ is then an $S^k \times S$ probability matrix where value ${}_k\alpha_{ij}$ models the probability of moving towards state $v_j \ \forall j \in \{1, 2, ..., S\}$ given the $i \in \{1, 2, ..., S^k\}$ k-length history. Figure 1a shows the probability matrix of such a 2nd order Markov chain, fitted on the entire DIALECT-2 dataset, where the rows represent the $5^2 = 25$ possibilities of a 2-length history given 5 blood glucose levels.

In higher order Markov chains, the main interest is still in the transition behaviour and under the assumption of stationary sequences, the k initial probability distributions are often ignored. If needed, those initial probabilities can be calculated by normalising over the last time point in the k-length history, just as we calculated π by normalising over all j target states. We denote the normalisation of ${}_k\mathbf{A}$ down to the ℓ^{th} order with a tilde: ${}^{\ell}_{k}\tilde{\mathbf{A}} \ \forall \ell \in \{0, 1, 2, ..., k-1\}$. Its parameters are then written as ${}^{\ell}_{k}\tilde{\alpha}_{ij}$. For this reason, the number of free parameters in a kth order Markov chain is ${}_kK = S^k(S-1)$.

In Sect. 4, we will discuss quality measures based on information-theoretic scoring functions. These quality measures use log likelihood to quantify the goodness of fit of a given Markov chain. Here, we start making a distinction between two datasets:

on the one hand, the dataset *for* which we calculate the goodness of fit, and on the other hand, the dataset *on* which we estimate the parameters of the Markov chain. We denote the latter with a superscript; ${}_k\mathbf{A}^F$ refers to a k$^{\text{th}}$ order transition matrix estimated on dataset F. We can now calculate the log likelihood of dataset G using parameter estimates ${}_k\mathbf{A}^F$ by

$$\mathcal{L}(P(G|{}_k\mathbf{A}^F)) = \sum_{h=1}^{S} n_h^{G^{(t=1)}} \log {}_k^0\tilde{\alpha}_h^F + \sum_{i=1}^{S}\sum_{j=1}^{S} n_{ij}^{G^{(t=\{1,2\})}} \log {}_k^1\tilde{\alpha}_{ij}^F + \ldots$$
$$+ \sum_{i=1}^{S^{k-1}}\sum_{j=1}^{S} n_{ij}^{G^{(t=\{1,2,\ldots,k\})}} \log {}_k^{k-1}\tilde{\alpha}_{ij}^F + \sum_{i=1}^{S^k}\sum_{j=1}^{S} n_{ij}^G \log {}_k\alpha_{ij}^F.$$

In the rest of this paper we use $\mathcal{L}(P(SG|\mathbf{A}^{\Omega}))$ and $\mathcal{L}(P(SG|\mathbf{A}^{SG}))$ to denote the log likelihood score for a subgroup using parameter values estimated on the entire dataset and on the subgroup respectively.[1]

3 Related work

Lemmerich et al. (2016) introduced an EMM model class for 1$^{\text{st}}$ order Markov chains. They focus on finding subgroups of transitions and thus detect heterogeneity within sequences. We propose to extend this model class such that 1) we find subgroups of entire sequences and detect homogeneous sequences that are heterogeneous with respect to the other sequences, and 2) we allow for discovering subgroups that are best modelled with a different order Markov chain.

Since Lemmerich et al. (2016) considered the situation where subgroups follow the same 1$^{\text{st}}$ order model as the entire dataset, they proposed a parameter-based quality measure related to the *total variation distance* or Manhattan distance:

$$\omega_{tv}({}_1\mathbf{A}^{SG}, {}_1\mathbf{A}^{\Omega}) = \sum_{i=1}^{S}\left(\sum_{j=1}^{S} n_{ij}^{SG} \sum_{j=1}^{S}\left|{}_1a_{ij}^{SG} - {}_1a_{ij}^{\Omega}\right|\right). \tag{4}$$

The quality measure ω_{tv} can be extended to situations where the subgroup follows the same higher order Markov chain as the entire dataset, but it cannot be used in situations where the subgroup follows a different order model.

Song et al. (2015, 2016) propose what they call Model-Based Subgroup Discovery (MBSD) where the divergence between the target probability estimates and the true labels of an outcome variable is evaluated using Proper Scoring Rules (PSR) (Gneiting and Raftery 2007). We analyse sequential data without labels, but our evaluation measures are still related to those in Song et al. (2016) since the information-theoretic scoring function AIC is derived from the Kullback-Leibler divergence (Burnham and

[1] Note that while calculating the log likelihood, we use normalised probabilities for the first k time points. In general, in this paper we assume homogeneous and stationary sequences, except for Sect. 5.2, where we analyse non-stationary sequences.

Anderson 2004), which is associated with the logarithmic score as a PSR (Gneiting and Raftery 2007).

In fact, for outcome variables with a probability density distribution, Song (2017) defines a quality measure called *weighted divergence* where the information gain of the subgroup is calculated using log likelihood as the negative of the expected loss, or information content,

$$\varphi_{WD}(SG, \theta^{SG}, \theta^{\Omega}) = \mathcal{L}(P(SG|\theta^{SG})) - \mathcal{L}(P(SG|\theta^{\Omega})). \tag{5}$$

Although our quality measures based on information-theoretic scoring functions only differ from φ_{WD} by the addition of a penalty for model complexity, it is exactly this penalty that allows for discovering subgroups with a different order Markov chain. In Sect. 5 we will show the difference in performance between our quality measures and φ_{WD}.

Several papers proposed information-theoretic scoring functions as the basis of a quality measure. For tabular data, Lijffijt et al. (2018) seek exceptional location and spread in multiple real-valued targets, by means of a quality measure based on the information gain of subgroups that allows the user to incorporate prior knowledge in the process. For graph data, Deng et al. (2020) aim to identify pairs of subgraphs with exceptional connectivity by looking at the density between the subgraphs, also allowing for the incorporation of prior knowledge.

In the context of sequence data, Bueno et al. (2020) consider dynamic Bayesian networks as model class and use BIC to define a mismatch score between the subgroup and its complement. We look at Markov chains, and compare the subgroup to the entire dataset because this is conceptually easier to understand and computationally more efficient than comparing to the subgroup's complement.

Like Lemmerich et al. (2016); Song (2017); Bueno et al. (2020), we take an approach where candidate subgroups are evaluated using a bottom-up, heuristic search through the descriptive space. In comparison, Becker et al. (2017) take a top-down approach where the subgroups are hypothesised beforehand based on theory and evaluated using Bayes Factors. Kiseleva et al. (2013) hypothesise two groups of sequences based on descriptive information and distributional characteristics. The two groups are analysed with a Markov model and compared on their prediction accuracy.

A global approach to detecting (groups of) outliers is taken by Sadagopan and Li (2008), who calculate the log likelihood scores of individual sequences under a 1^{st} order Markov model. Specifically, they create a Mahalanobis distribution of sequences by combining these log likelihood scores with meta information such as the sequence length. Although their approach points at unusual sequences in an existing dataset, it does not describe or explain in any other way why specifically those sequences are considered outliers. In contrast, the framework of EMM is a local pattern mining technique that not only allows for interpretable descriptions of exceptional subgroups but also for an explanation of why those subgroups are selected.

Yet, in Sect. 5, we will compare our method to a quality measure that only uses a globally fitted model and does not require the fit of separate models in each candidate subgroup. In particular, we compare against a quality measure called *weighted relative*

likelihood (Song 2017),

$$\varphi_{\text{WRL}}(SG, \Omega, \theta^{\Omega}) = M^{SG} \cdot \left| \frac{\mathcal{L}(P(\Omega|\theta^{\Omega}))}{M^{\Omega}} - \frac{\mathcal{L}(P(SG|\theta^{\Omega}))}{M^{SG}} \right|. \tag{6}$$

Here, the average fit of the dataset is evaluated under parameters estimated on the dataset and compared against the average fit of the subgroup under those same parameters. Note that M^{SG} and M^{Ω} denote the total number of observations in the subgroup and dataset respectively and not the total number of sequences (see Sect. 2.1).

Recently, Mollenhauer and Atzmueller (2020) propose Sequential Exceptional Pattern Discovery using Pattern-Growth (*SEPP*) as a general approach to the problem of Sequential Exceptional Model Mining (SEMM). Their work combines EMM with sequential pattern mining, the task to identify frequent subsequences, and as such they develop a search strategy based on GP-growth (Lemmerich et al. 2012) and PrefixSpan (Pei et al. 2004). Mathonat et al. (2021) also combine sequential pattern mining with EMM, developing the *MCTSExtent* method building on Monte Carlo Tree Search (MCTS) (Bosc et al. 2018). Both MCTSExtent and SEPP consider the data to be in the traditional sequential pattern mining form where X_t is an itemset. Such a dataset is inherently binary: an item is present in an itemset or not. A subgroup's exceptionality is then evaluated in terms of frequency or precision. In contrast, our sequences come with m descriptive attributes; subgroups are formed using these attributes and evaluated based on exceptional sequential behaviour.

4 Proposed method: Exceptional transition behaviour of any order

We will now explain how we derive our quality measures based on information-theoretic scoring functions in Sect. 4.1. In Sect. 4.2, we then discuss the proposed search strategy.

4.1 Quality measures based on information-theoretic scoring functions

In order to evaluate the exceptionality of candidate subgroups with varying order Markov chains, we develop a set of quality measures that allows for the comparison of two sets of parameters of different size. Such quality measures should not simply select the subgroup with the largest number of parameters, because a more complex model may overfit the data. The quality measures should further take the subgroup size into account, since deviations from the norm are more easily obtained in smaller subgroups.

To that end, we base our quality measures on information-theoretic scoring functions. In general, for a dataset G, such as scoring function is defined by

$$\phi_{LL}(G, \theta^{G}) = \mathcal{L}(P(G|\theta^{G})) - f(M^{G}) \cdot K^{G}, \tag{7}$$

where we use subscript LL to indicate that we use log likelihood as a way to quantify the goodness of fit. Given that θ^{G} are MLE parameters, $\mathcal{L}(P(G|\theta^{G})$ is maximal. The

second part of the equation is a penalty for model complexity. Here, K^G denotes the number of free parameters in a model estimated on dataset G. The term $f(M^G)$ is a penalty based on the number of observations in G.

We will apply three information-theoretic scoring functions which all use a different penalty term. First, Akaike (1973, 1974) (see also Burnham and Anderson 2004) derived that the bias in the log likelihood score (due to overfitting) converges to K as $M \to \infty$ (we temporarily leave out the superscript G). In Akaike's Information Criterion (AIC), the penalty is therefore set to K, which means that $f(M) = 1$. Second, the Bayesian Information Criterion (BIC) (Schwarz 1978) sets $f(M) = \frac{1}{2} \log M$. In this paper, we refer to BIC as an information-theoretic scoring function because it only differs from AIC by the extent of the penalty. However, BIC is derived from a Bayesian viewpoint, is related to Bayes factors (Kass and Raftery 1995) and originally focuses on model selection instead of prediction accuracy (Pohle et al. 2017). Third, a scoring function called AIC with small sample correction (AICc) penalises with an additional $\frac{2K^2+2K}{M-K-1}$. The term corrects for overfitting if the number of free parameters is large with respect to M, but AICc converges to AIC as M increases (Sugiura 1978; Hurvich and Tsai 1995; Burnham and Anderson 2004).

Several authors have investigated the use of AIC and BIC in determining the appropriate Markov chain order (e.g. Tong 1975; Schoof and Pryor 2008; Singer et al. 2014). We will now use these scoring functions to evaluate whether candidate subgroups have exceptional transition behaviour. This goes as follows.

In the situation that dataset Ω is heterogeneous and contains one or more subgroups of sequences that follow a different model than the rest of the sequences, it is likely that the parameters of the subgroup, θ^{SG}, describe the subgroup better than the parameters of the entire dataset, θ^{Ω}. This means that in the presence of a subgroup, the log likelihood of dataset Ω will increase if the parameters of the subgroup are separately estimated and evaluated. We write that

$$\mathcal{L}(P(SG|\theta^{SG})) + \mathcal{L}(P(SG^C|\theta^{\Omega})) > \mathcal{L}(P(\Omega|\theta^{\Omega})), \tag{8}$$

where SG^C denotes the subgroup's complement. Since $\mathcal{L}(P(SG^C|\theta^{\Omega}))$ is part of both the left and the right side of Eq. (8), we can write that

$$\mathcal{L}(P(SG|\theta^{SG})) > \mathcal{L}(P(SG|\theta^{\Omega})). \tag{9}$$

We now derive our quality measures by combining Eqs. (7) and (9). We furthermore multiply ϕ_{LL} with -2 for conventional reasons, and again multiply with -1 to obtain quality measures that should be maximised (see Definition 2). This gives us the following three quality measures.

$$\varphi_{AIC} = 2\mathcal{L}(P(SG|\theta^{SG})) - 2K^{SG} - 2\mathcal{L}(P(SG|\theta^{\Omega})) + 2K^{\Omega}, \tag{10}$$

$$\varphi_{BIC} = 2\mathcal{L}(P(SG|\theta^{SG})) - K^{SG} \log M^{SG} - 2\mathcal{L}(P(SG|\theta^{\Omega})) + K^{\Omega} \log M^{SG}, \tag{11}$$

$$\varphi_{AICc} = 2\mathcal{L}(P(SG|\theta^{SG})) - 2K^{SG} - \frac{2K^{SG^2} + 2K^{SG}}{M^{SG} - K^{SG} - 1} -$$

$$2\mathcal{L}(P(SG|\theta^{\Omega})) + 2K^{\Omega} + \frac{2K^{\Omega^2} + 2K^{\Omega}}{M^{SG} - K^{\Omega} - 1}. \tag{12}$$

Note that quality measure φ_{WD} as defined earlier in Eq. (5) sets $f(M) = 0$ and therefore uses no penalty. Furthermore, if the subgroup has the same order Markov chain model as the dataset, $K^{\Omega} = K^{SG}$ and the penalty terms cancel out for all three quality measures.

4.2 Extended beam search algorithm

Beam search is a commonly used strategy to search through the space of candidate subgroups. It has the ability to use descriptive attributes from any domain (i.e. it can natively handle any mix of attributes that are binary, categorical or numerical without the requirement for static pre-algorithm discretisation). The algorithm (Duivesteijn et al. 2016, Algorithm 1, page 60) performs a level-wise search of d levels, where at each level the descriptions of a set of candidate subgroups are further refined and evaluated with a quality measure. The w best-scoring subgroups are selected for the next level. In the end, the algorithm outputs a list of the top-q subgroups.

In order to find subgroups with varying order Markov chains, we have to take a few additional steps to evaluate candidate subgroups. First, we have to find the Markov chain order that best fits the entire dataset Ω. Algorithm 1 describes the procedure. As explained in Sect. 2.3, a higher order Markov chain can be transformed into a lower order model by normalising the transition matrix. We use this feature to calculate the transition probabilities only once using a Markov chain of order s (line 2), where s is a user-defined parameter which we call the *start* parameter. In line 3, we calculate the penalised log likelihood given penalty $p \in \{AIC, BIC, AICc\}$ as described in Sect. 4.1. In lines 5-13, we repeatedly normalise the transition matrix (line 7), calculate the new score (line 8) and check whether the score has increased or not (line 9). The

Algorithm 1 Finding the best fitting Markov chain order

Input A dataset G, a penalty p from $\{AIC, BIC, AICc\}$, start parameter s
Output The estimated Markov chain parameters, the Markov chain order

1: **procedure** BESTFITTINGORDER
2: $\quad {}_s\mathbf{A}^G \leftarrow$ MARKOVCHAIN(G, order $= s$)
3: $\quad \text{score}_s \leftarrow \phi_{LL}(G, {}_s\mathbf{A}^G, p)$ ▷ Eq. (7), with penalty term replaced by p
4: $\quad$ counter = 1
5: $\quad$ **while** $f < s$ **do**
6: $\quad\quad \ell = s - f$
7: $\quad\quad {}^{\ell}_{s}\tilde{\mathbf{A}}^G \leftarrow \text{normalise}({}_s\mathbf{A}^G)$
8: $\quad\quad \text{score}_{\ell} \leftarrow \phi_{LL}(G, {}^{\ell}_{s}\tilde{\mathbf{A}}^G, p)$ ▷ Eq. (7), with penalty term replaced by p
9: $\quad\quad$ **if** $\text{score}_{\ell} < \text{score}_s$ **then**
10: $\quad\quad\quad$ **return** ${}^{\ell+1}_{s}\tilde{\mathbf{A}}^G, \ell + 1$
11: $\quad\quad$ **else**
12: $\quad\quad\quad$ counter = counter + 1
13: $\quad\quad\quad \text{score}_s = \text{score}_{\ell}$
14: $\quad$ **return** ${}^{\ell}_{s}\tilde{\mathbf{A}}^G, \ell$

Algorithm 2 Evaluating a subgroup with varying order Markov chains

Input A subgroup SG, a penalty p from $\{AIC, BIC, AICc\}$ and according quality measure φ, dataset parameters ${}^{\ell}_{s}\tilde{\mathbf{A}}^{\Omega}$, start parameter s

Output Real number expressing the exceptionality of subgroup SG

1: **procedure** EVALUATINGSUBGROUP
2: ${}^{u}_{s}\tilde{\mathbf{A}}^{SG}, u \leftarrow$ BESTFITTINGORDER(SG, p, s)
3: quality $\leftarrow \varphi(SG, {}^{u}_{s}\tilde{\mathbf{A}}^{SG}, {}^{\ell}_{s}\tilde{\mathbf{A}}^{\Omega})$
4: **return** quality

procedure returns the parameter estimates and the Markov chain order of the model that maximises the penalised log likelihood fit.

Algorithm 2 describes how a candidate subgroup is evaluated. First, procedure BESTFITTINGORDER is repeated for the subgroup (line 2). Then, the subgroup is evaluated with quality measure φ_{AIC}, φ_{BIC} or φ_{AICc} (Eqs. (10), (11) and (12) respectively), depending on parameter $p \in \{AIC, BIC, AICc\}$. Like usual, beam search returns the q best scoring subgroups.

The worst-case computational complexity of the beam search algorithm is $\mathcal{O}(dwZE(c + \mathcal{M}(N, f) + \log(wq)))$ (Duivesteijn et al. 2016, page 60, we have slightly adapted the notation to avoid confusion). Here, d, w, q are as explained earlier, Z is the number of descriptors and E the worst-case number of nominal values (numerical and binary descriptors are refined faster). Parameter c refers to the complexity of comparing two models. In our approach, we compare the fit of the subgroup under two models (θ^{SG} and θ^{Ω}) but the parameters of the data model are calculated only once at the beginning of the beam search. The term $\mathcal{M}(N, f)$ refers to the cost of learning a model $\mathcal{M}$ from N records on f targets. In case of Markov chains, this would compare to fitting a k$^{\text{th}}$ order model for S state-values on N sequences of length T. The computational complexity is then a linear function of N, T and the number of free parameters K, which grows exponentially with base S and exponent s (Sect. 2.3).

The fact that we can evaluate lower order Markov chains by normalising higher order transition matrices is a powerful feature that keeps the computational complexity of our approach tractable. Still, parameter s is an important parameter because if s is too large, fitting the Markov chain model may take unnecessarily long, while if s is too small, it is hard to evaluate more complex models. Note furthermore that s determines which parts of the sequences are used for model fitting. After all, fitting a k$^{\text{th}}$ order Markov chain can only be done with the data from time points $k + 1$ to T. Although it is possible to normalise higher order transition matrices all the way down to a 1$^{\text{st}}$ order Markov model, the drawback of such a procedure is that not all observations are used for estimating the probabilities.

5 Experiments on synthetic data

In the following, we assess the performance of our proposed method by means of experiments on synthetic data. The goal of the experiments is to see whether the proposed quality measures can indeed detect subgroups with exceptional transition

behaviour of varying order. For varying data characteristics such as sequence length and the state-space, we will create ground truth subgroups and analyse whether they are ranked first in the top-q result list. Specifically, Sect. 5.1 analyses exceptional transition behaviour, Sect. 5.2 analyses exceptional starting behaviour and Sect. 5.3 contains a sensitivity analysis.

5.1 Exceptional transition behaviour of varying order

5.1.1 Experimental methodology

We generate synthetic data with $N = 100$ sequences with $T \in \{10, 50, 200\}$ time points, $S \in \{2, 5, 10\}$ states and $Z \in \{5, 10, 20\}$ binary, descriptive attributes.[2] The descriptive attributes are sequence-level attributes, as explained in Sect. 2. For each sequence, $p(a_z = 1) = 0.5$ for $z \in \{1, 2, ..., Z\}$. A ground truth subgroup is defined for sequences where $a_1 = 1 \wedge a_2 = 1$. Thus, approximately 25% of the sequences are part of the true subgroup. All other sequences follow a 1st order Markov chain with probabilities drawn from a uniform probability distribution. The probabilities are normalised to sum to 1. In line with the assumption of stationary sequences, the first time points are sampled using a normalisation as well.

Two types of subgroups are generated, as specified by simulation parameter $order \in \{1, 2, 3, 4\}$.

1. If order $= 1$, the subgroup has an exceptional 1st order transition model. This means that both the subgroup and the rest of the dataset follow a 1st order Markov chain, but the parameter values of the subgroup are different. We can write that ${}_1\mathbf{A}^{SG} \neq {}_1\mathbf{A}^{\Omega}$.
2. If order $= k$ for $k \in \{2, 3, 4\}$, the subgroup follows a kth order Markov model. This means that the subgroup is best modelled with $S^k \times S$ transition matrix ${}_k\mathbf{A}^{SG}$ while the rest of the data is modelled with a 1st order transition model ${}_1\mathbf{A}^{\Omega}$. Here, Algorithm 1 will fit a 1st order Markov chain to the entire dataset, and the subgroups should be fitted with a more complex model.

Every combination of simulation parameters is repeated $nreps = 50$ times.

Given a synthetic dataset with a ground truth subgroup, we perform EMM with 6 different quality measures. First, we apply the three quality measures based on information-theoretic scoring functions as proposed in Sect. 4.1: φ_{AIC}, φ_{BIC} and φ_{AICc}. We compare our quality measures against three reference measures as mentioned in Sect. 3: ω_{tv}, φ_{WD} and φ_{WRL} (Eqs. (4), (5) and (6) respectively).

Since ω_{tv} is a parameter-based quality measure, we cannot use it to evaluate subgroups of varying order. Instead, we will first determine the Markov chain order of the entire dataset by applying Algorithm 1 with $p = AIC$ and then evaluate candidate subgroups using that same order. Similarly for φ_{WRL}. In case of φ_{WD}, $p = AIC$ when determining the Markov chain order of the dataset, but candidate subgroups are evaluated with $p = none$. Note that since 75% of the sequences are generated with a 1st order Markov chain, it is unlikely that we will evaluate subgroups against higher

[2] Source code available at github.com/RianneSchouten/simulations_markov_chains_emm.

order models. However, determining the order of the entire dataset is an important step when analysing real-world data (see Sect. 6).

For each quality measure, we save the rank of the ground truth subgroup in the $q = 20$ output of the extended beam search algorithm. Since every dataset undoubtedly contains 1 ground truth subgroup, we expect the quality measures to give a first rank to that subgroup. We furthermore check the estimated Markov chain order of the ground truth subgroup and calculate the percentage of simulation repetitions where the correct order is given. If the result list does not contain the ground truth subgroup, we set the rank to $q + 1$ and the order to NaN.

The other parameters of the beam search are $w = 25$ and $d = 3$. We furthermore constrain the subgroup to minimally contain 10% of the sequences. We set start parameter $s = 4$.

5.1.2 Results

Figures 2 and 3 respectively show the rank and the order of the ground truth subgroup. We present the results for $Z = 20$ descriptive attributes. Our information-theoretic based quality measures φ_{BIC}, φ_{AIC}, and φ_{AICc} give similar results and they are therefore presented in one row. Clearly, they give a first rank to the ground truth subgroup when the sequences are relatively long ($T \geq 50$). For shorter sequences ($T = 10$), the subgroup is sometimes, but not always, ranked first. Here, it is easier to find the ground truth subgroup if the state-space is small, the ground truth order is close to 1 and the descriptive space is small (the latter is not visible in Fig. 2).

Doubtlessly, the ground truth order of the subgroup can only be detected if there are enough observations. First, the estimation of a k^{th} order Markov chain requires $T > k$ time points, and $M > S^k(S-1)$ observations. Second, a larger subgroup allows for a more precise estimation of the Markov chain. For instance, for a state-space of 2, a subgroup with a 1^{st}, 2^{nd}, 3^{rd}, or 4^{th} order Markov chain can be found when the sequences are long ($T = 200$, Fig. 3). However, when the state-space increases or the number of time points decreases, it is not always possible to detect higher order Markov models (the subgroups are still ranked first, though). Note that in the experiment, we fixed the number of sequences to $N = 100$. In addition to increasing the sequence length, increasing the number of sequences would possibly also allow for a correct estimation of the Markov chain order.

Quality measures φ_{AIC} and φ_{AICc} are slightly more robust for the number of observations than φ_{BIC}. We can see this in Fig. 3, where φ_{AIC} finds subgroups with a $k = 2$ Markov chain order when $S = 5$ and $T = 50$ or $S = 10$ and $T = 200$, or with a $k = 3$ order when $S = 5$ and $T = 200$. In contrast, φ_{BIC} finds the order in none of those subgroups. These findings seem logical since we know that the BIC uses a larger penalty than AIC (Sect. 4.1). We do not see important differences between φ_{AIC} and φ_{AICc}.

We know that φ_{WD} does not use a penalty. Therefore, the estimated order will always be equal to start parameter s, since a more complex model will have a better log likelihood fit. The only limitation is $M < K$, which happens when, for instance, $T = 50$, $S = 10$, and $s \in \{3, 4\}$. Consequently, in Fig. 3, 100% of the subgroups

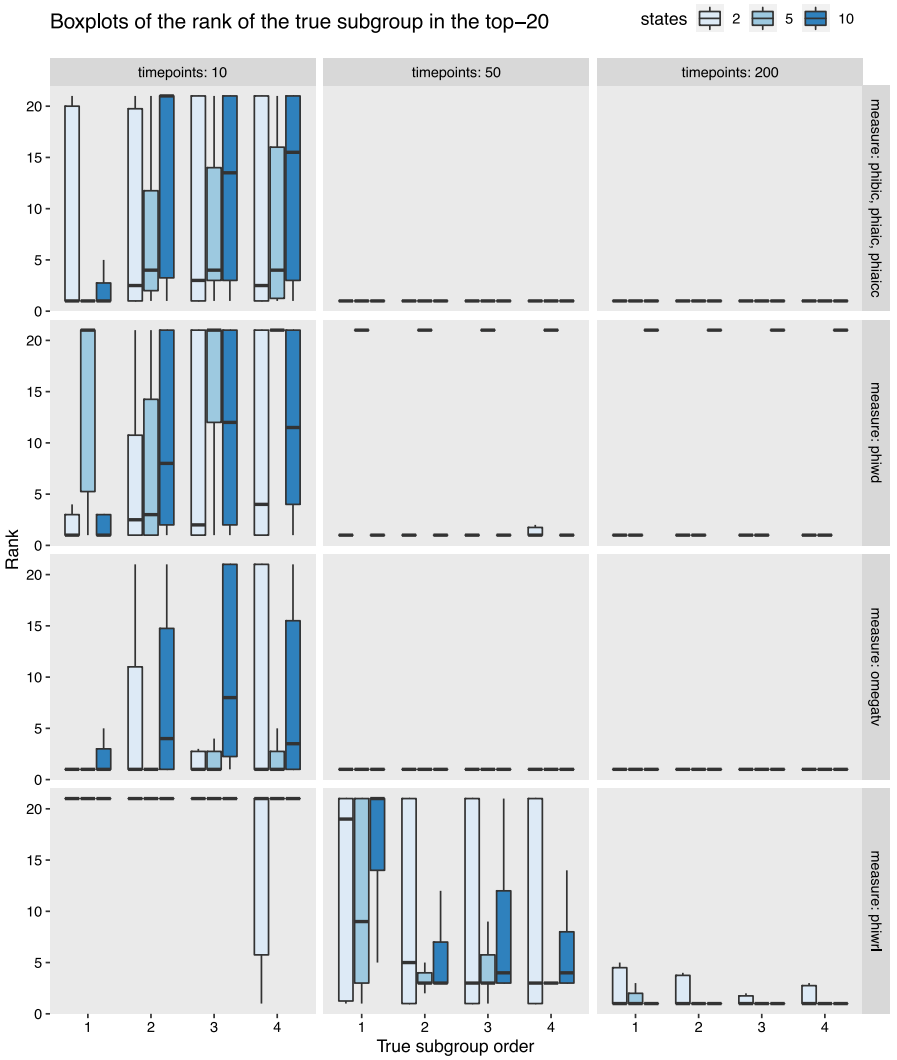

Fig. 2 Boxplots of the rank of the ground truth subgroup. The ideal value is a rank of 1. We present the simulation results for 20 descriptive attributes ($Z = 20$) and 50 repetitions ($nreps = 50$). Results for φ_{BIC}, φ_{AIC}, and φ_{AICc} are similar and therefore presented in one row

with a true 2nd order Markov model are found, but none of the subgroups with a true Markov model of other orders. Similar results are obtained when $S = 5$ and $T = 200$.

Although φ_{WD} fits the wrong 2nd order Markov chain to all subgroups when $T = 50$ and $S = 10$, almost all subgroups are still ranked first (Fig. 2). By contrast, when $T = 50$ and $S = 5$, subgroups are also estimated with the wrong model (3rd order), but these do not end up in the top-20 result list. We obtain similar results for $S = 10$ and $T = 200$ (Fig. 2). Apparently, for $S = 10$, the availability of longer sequences amplifies the difference between the subgroup and the entire dataset, and estimating

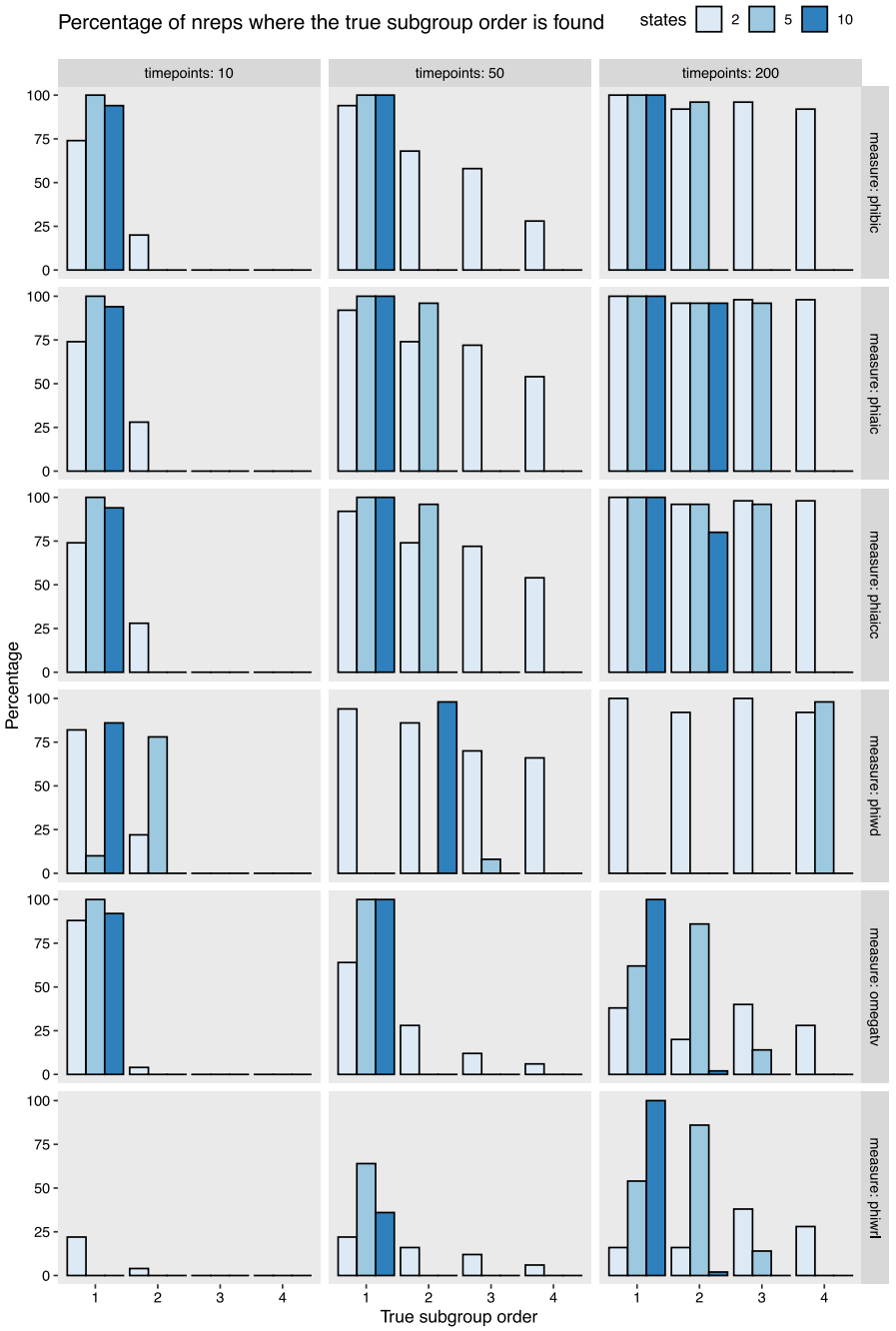

Fig. 3 Percentage of the number of simulations where the true order of the subgroup is found. The ideal value is 100%. We present the simulation results for 20 descriptive attributes ($Z = 20$) and 50 repetitions ($nreps = 50$)

the wrong order therefore disturbs the ranking, while for $S = 5$, the availability of longer sequences allows for a low ranking even though the estimated Markov chain order is wrong.

The last two rows in Figs. 2 and 3 show the results for ω_{tv} and φ_{WRL}. Both measures evaluate subgroups using the order as estimated on the entire dataset. When $T = 200$, the estimated dataset order sometimes equals the subgroup order, and ω_{tv} and φ_{WRL} will therefore sometimes correctly estimate the ground truth Markov chain order (Fig. 3).

It may still be surprising that ω_{tv} and φ_{WRL} give a first rank to ground truth subgroups with a higher order Markov chain model (Fig. 2). The likely reason for ω_{tv} is as follows. As discussed in Sect. 4.1, the difference between a normalised and a directly estimated 1st order transition model is that the latter uses the observations of all available time points whereas the first uses only time points $k + 1$ to T. For long sequences, this difference is negligible and therefore ω_{tv} (which uses all time points) finds parameter estimates that are close to reality. However, for short sequences with a large number of states, noise disturbs the estimation.

Quality measure φ_{WRL} greatly relies on the parameters that are estimated on the entire dataset. We see that the higher the ground truth order, the more frequently the ground truth subgroup is ranked first ($T = 50$, Fig. 3). Possibly, when the subgroup follows a higher order Markov model, the dataset parameters are more directed towards the subgroup's complement than when the subgroup follows a 1st order Markov model.

5.2 Exceptional starting behaviour

5.2.1 Experimental methodology

We further evaluate subgroups of sequences with exceptional *initial* probabilities, or exceptional starting behaviour. In other words, the subgroup follows the same 1st order transition model with the same parameter values as the rest of the data: ${}_1\mathbf{A}^{SG} = {}_1\mathbf{A}^{\Omega}$. However, the subgroup has a distinct set of initial probabilities (Eq. (3)), which should not be modelled with normalised probabilities but with a separate set of probability values: $\pi^{SG} \neq {}^{0}_{1}\tilde{\mathbf{A}}^{SG}$. We thus reject the assumption of stationary sequences (see Sect. 2.3).

Here, the number of free parameters in the subgroup is $S(S-1)$ for the transition probabilities and an additional $S-1$ free parameters for the initial probabilities vector π. In practice, such a model only makes sense when sequences are very short. We therefore decide to run the simulation for exceptional starting behaviour with parameters $N \in \{100, 500, 1000\}$, $S \in \{2, 5, 10\}$, $T \in \{2, 5, 10\}$, $Z \in \{5, 10, 20\}$, $s = 1$ and $nreps = 25$. Again, the beam search parameters are $q = 20$, $w = 25$, $d = 3$ and a subgroup should cover at least 10% of all sequences.

5.2.2 Results

Subgroups with exceptional starting behaviour follow the same 1st order transition model as the rest of the dataset, but have a distinct pattern for the very first time point.

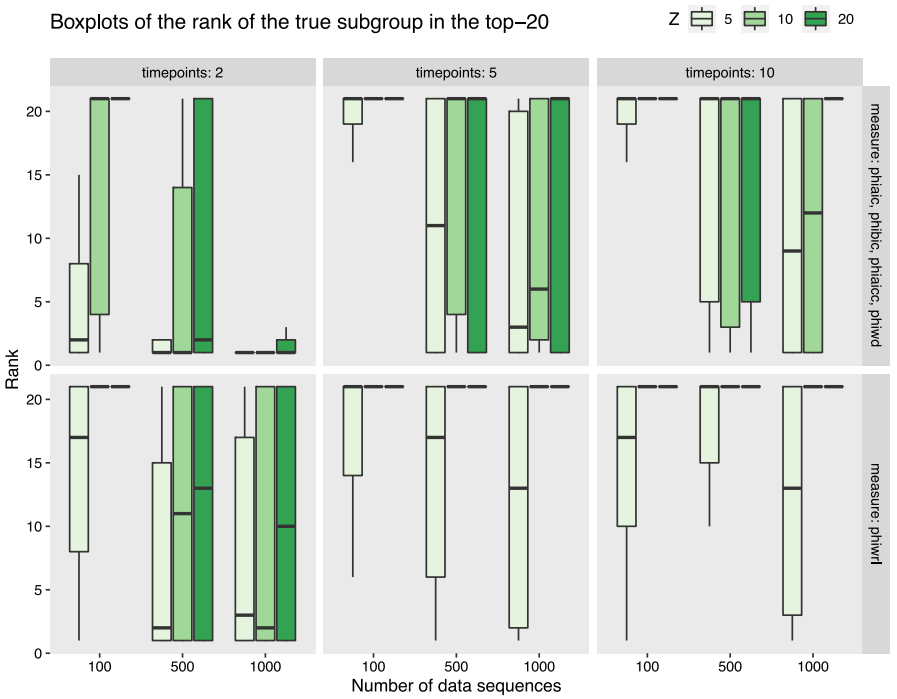

Fig. 4 Boxplots of the rank of the ground truth subgroup with exceptional starting behaviour. The ground truth subgroup differs from the rest of the dataset by its initial probabilities. The transition behaviour of subgroup and dataset is the same. The ideal value is a rank of 1. We present the simulation results for 5 states ($S = 5$) and 25 repetitions ($nreps = 25$). Results for φ_{BIC}, φ_{AIC}, φ_{AICc}, and φ_{WD} are similar and therefore presented in one row. Quality measure ω_{tv} cannot detect exceptional starting behaviour and is therefore not shown

Figures 4 and 5 present our findings for a state-space of 5. In general, the smaller the state-space, the more advantageous the result.

It turns out that the log likelihood based quality measures (either with or without penalty) perform comparably in ranking the ground truth subgroup. Therefore, these measures are shown in a single row in Fig. 4. These quality measures give a first rank to the ground truth subgroup when (1) there are enough sequences, (2) the sequences are not too long and (3) there are not too many descriptive attributes. Although it is difficult for evaluation measures φ_{AIC}, φ_{BIC} and φ_{AICc} to give a first rank to a ground truth subgroup with exceptional starting behaviour, especially if the sequences are long, these information-theoretic scoring functions do allow for a correct estimation of the Markov chain order (see Fig. 5). The reason is that exceptional starting behaviour causes an increase in the number of free parameters (see Sect. 2.3). As a result, for short sequences ($T = 2$), the penalties are too large to counter-effect the increase in log likelihood. For long sequences however, the model fit increases sufficiently. Logically, since φ_{WD} does not use a penalty, it is good at estimating more complex models (Fig. 5).

Both ω_{tv} and φ_{WRL} evaluate candidate subgroups using the same Markov chain order as estimated on the entire dataset, which is a 1st order chain without additional

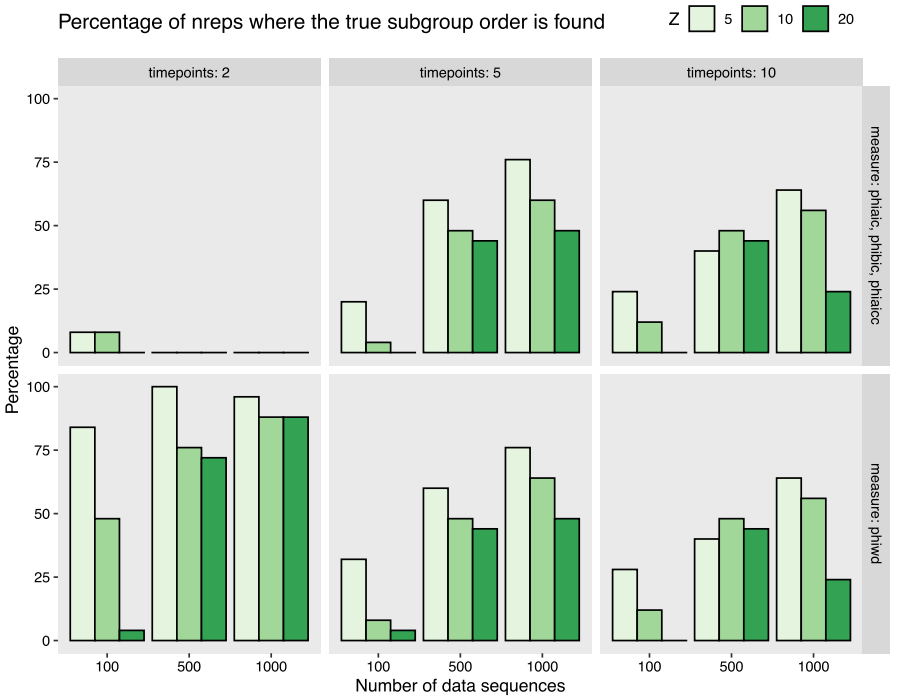

Fig. 5 Percentage of the number of simulations where the ground truth subgroup is found. The ground truth subgroup differs from the rest of the dataset only by its initial probabilities: transition behaviour of subgroup and dataset is the same. The ideal percentage is 100%. We present the simulation results for 5 states ($S = 5$) and 25 repetitions ($nreps = 25$). Results for φ_{BIC}, φ_{AIC}, and φ_{AICc} are similar and therefore presented in one row. Quality measures ω_{tv} and φ_{WRL} cannot find the ground truth subgroup order and are therefore not shown

initial parameters. Using such a model, ω_{tv} never manages to rank the true subgroup first (and we therefore omit the results from the figures), while φ_{WRL} achieves it sometimes when the sequences are as short as possible ($T = 2$) and there are only $Z = 5$ descriptive attributes (Fig. 4).

5.3 Sensitivity analysis

In Sect. 5.1, we analysed the performance of the quality measures for the setting where the global model is fitted with a 1st order Markov chain, the start parameter $s = 4$, and the subgroups are fitted with a Markov chain order between 1 and 4. The combination of these settings allows our algorithms to find the correct order. In Sect. 5.3.1, we ask ourselves what would happen if the parameters are misspecified such that the algorithms are steered away from finding the correct order. In Sect. 5.3.2, we vary the subgroup size and description length.

Table 1 Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{AIC}

		$T = 10$		$T = 50$		$T = 200$	
		True SG order		True SG order		True SG order	
Gl.order	Start s	2	3	2	3	2	3
1	2	2 (16)	3 (3)	1 (0)	1 (0)	1 (0)	1 (0)
	4	11 (17)	13 (20)	1 (0)	1 (0)	1 (0)	1 (0)
3	2	21 (0)	21 (0)	1 (0)	21 (11)	1 (0)	1 (0)
	4	21 (0)	21 (0)	1 (0)	21 (0)	1 (0)	1 (0)

The order of the global model is either 1 or 3, and the search is started with parameter $s = 2$ or $s = 4$. We show the results for subgroups where $order \in \{2, 3\}$ and for sequences with length $T \in \{10, 50, 200\}$. Further, $N = 100$, $Z = 20$, $S = 5$, and $nreps = 10$

Table 2 Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{WD}

		$T = 10$		$T = 50$		$T = 200$	
		True SG order		True SG order		True SG order	
Gl.order	Start s	2	3	2	3	2	3
1	2	4 (12)	21 (4)	1 (0)	1 (0)	1 (0)	1 (0)
	4	5 (19)	21 (0)	21 (0)	21 (0)	1 (0)	1 (0)
3	2	21 (0)	21 (0)	1 (0)	21 (0)	1 (0)	1 (0)
	4	21 (0)	21 (0)	21 (0)	21 (0)	1 (0)	1 (0)

The order of the global model is either 1 or 3, and the search is started with parameter $s = 2$ or $s = 4$. We show the results for subgroups where $order \in \{2, 3\}$ and for sequences with length $T \in \{10, 50, 200\}$. Further, $N = 100$, $Z = 20$, $S = 5$, and $nreps = 10$

5.3.1 Varying global model order and varying start parameter *s*

First, we investigate the effect of:

1. Changing the start parameter to $s = 2$
2. Changing the global model to a 3rd order Markov chain.

Therefore, we sample $N = 100$ sequences with a state-space of $S = 5$ and $Z = 20$ descriptive attributes. We vary the length of the sequences with $T \in \{10, 50, 200\}$. The simulation is repeated $nreps = 10$ times. The beam search settings are as before.

Tables 1 and 2 show the median and interquartile range (IQR) of the rank of the ground truth subgroup for quality measures φ_{AIC} and φ_{WD} respectively, for subgroups with $order \in \{2, 3\}$. We first inspect the results for φ_{AIC} for sequences where $T = 10$. It is clear that when the global model is fitted with a 1st order Markov chain, it is advantageous to set the start parameter to $s = 2$ instead of $s = 4$. In Table 1, we see that the median rank decreases from 11 (13) to 2 (3) for subgroups with a 2nd (3rd) order Markov chain. It is surprising that we can give a high rank to 3rd order subgroups using start parameter $s = 2$. Consistent with earlier findings, apparently it

can happen that subgroups are considered exceptional even when their Markov chain order is wrongly estimated. Note that these findings hold when T increases.

When the global model is a 3rd order Markov chain and we start evaluating at $s = 2$, the parameter settings forbid the algorithm to correctly estimate the parameters of the global model. However, we see that when T is sufficiently large, φ_{AIC} still ranks the ground truth subgroup first ($T = 200$, Table 1). On the other hand, when $T = 50$, it is difficult to find subgroups with a 3rd order Markov chain (but considering the IQR of 11 when $s = 2$, some subgroups can still be found).

For φ_{WD} and a global model order of 1, starting at $s = 2$ instead of $s = 4$ does not decrease the median rank, but it does positively affect the interquartile range ($T = 10$, Table 2). When the order of the global model increases from 1 to 3, φ_{WD} allows for discovering ground truth subgroups of order 2 when $s = 2$ and $T = 50$. However, when $T \leq 50$, these subgroups cannot be found using $s = 4$ (and neither can subgroups with $order = 3$). When $T = 200$, all subgroups can be found (just as was the case for φ_{AIC}).

The results for the other quality measures are not shown here but can be accessed in our repository.[3] In sum, quality measures φ_{BIC} and φ_{AICc} perform similarly to φ_{AIC} (Table 1), although the IQR of φ_{BIC} is sometimes a bit larger, especially when the global model order is 3 and $T = 200$. For φ_{WRL}, setting $s = 2$ instead of $s = 4$ is advantageous but only when the global model is a 1st order Markov chain. When the order of the global model is 3, φ_{WRL} has trouble finding the ground truth subgroup. Even though ω_{tv}, like φ_{WRL}, greatly depends on the estimated order of the global model, ω_{tv} has large IQRs. This indicates that even when the subgroup order is wrongly estimated, the subgroup can still be found. We saw similar results in Fig. 2.

Altogether, for shorter sequences, it can be advantageous to decrease the start parameter s. This applies both to 1st and 3rd order global models. In addition, when the global model has a 3rd order Markov chain, the ground truth subgroup can still be found as long as there are enough observations. This holds even when the starting parameter is set to two, which forbids the algorithm from considering the correct order.

5.3.2 Varying subgroup size and varying description length

Second, we investigate the effect of subgroup size and description length on the performance of our quality measures. Therefore, we vary:

1. The description length with $L \in \{1, 2\}$
2. The probability $p(a_z = 1) = pr$ with $pr \in \{0.35, 0.5\}$ for $z \in \{1, 2, ..., Z\}$.

Note that in Sect. 5.1, $pr = 0.5$ and $L = 2$, resulting in a subgroup that contains 25% of all sequences. Here, we will evaluate subgroups with a coverage of 13% ($pr = 0.35, L = 2$), 25% ($pr = 0.5, L = 2$), 35% ($pr = 0.35, L = 1$) and 50% ($pr = 0.5, L = 1$). We vary the number of descriptive attributes with $Z \in \{5, 10, 20\}$, set parameters $N = 100$, $S = 5$, and $T = 50$, and we model the global model with a 1st order Markov chain. Like before, we start our search at $s = 4$ and set $q = 20$, $w = 25, d = 3$, and the minimum subgroup size to 10%. This means that theoretically all subgroups could be found. We run the simulation $nreps = 10$ times.

[3] All results available at github.com/RianneSchouten/simulations_markov_chains_emm.

Table 3 Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{AIC}

		$Z = 5$		$Z = 10$		$Z = 20$	
		True SG order		True SG order		True SG order	
Desc.length	Prob.	1	4	1	4	1	4
1	0.5	2 (1)	2 (1)	1 (1)	2 (0)	2 (1)	1 (0)
	0.35	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (1)
2	0.5	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	0.35	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)

The subgroup description has $L \in \{1, 2\}$ attributes, where the probability per attribute is $pr \in \{0.35, 0.5\}$. We furthermore vary the number of descriptors $Z \in \{5, 10, 20\}$ and set $N = 100$, $T = 50$, $S = 5$ and $nreps = 10$. Results are presented for subgroups with a true Markov chain $order \in \{1, 4\}$

Table 4 Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{WRL}

		$Z = 5$		$Z = 10$		$Z = 20$	
		True SG order		True SG order		True SG order	
Desc.length	Prob.	1	4	1	4	1	4
1	0.5	11 (14)	1 (0)	12 (20)	1 (0)	2 (20)	1 (0)
	0.35	3 (11)	1 (0)	12 (12)	1 (0)	18 (18)	2 (3)
2	0.5	3 (0)	3 (0)	3 (10)	3 (1)	3 (1)	3 (1)
	0.35	15 (9)	8 (6)	15 (9)	10 (9)	21 (0)	18 (15)

The subgroup description has $L \in \{1, 2\}$ attributes, where the probability per attribute is $pr \in \{0.35, 0.5\}$. We furthermore vary the number of descriptors $Z \in \{5, 10, 20\}$ and set $N = 100$, $T = 50$, $S = 5$ and $nreps = 10$. Results are presented for subgroups with a true Markov chain $order \in \{1, 4\}$

Tables 3 and 4 present the results for quality measures φ_{AIC} and φ_{WRL} respectively, for subgroups with $order \in \{1, 4\}$. Quality measures φ_{BIC} and φ_{AICc} give similar results as in Table 1. There, we see that in almost all simulation settings, φ_{AIC} gives a first rank to all subgroups (with an IQR of 0). When $L = 2$ and $pr = 0.5$, it is a bit harder to find the true subgroup, although it is often still ranked second (Table 3). Clearly, when the true subgroup is large, it is more difficult to distinguish the subgroup from its complement.

For φ_{WRL}, we have already seen that it can find subgroups with a higher order Markov chain (cf. Sect. 5.1). In Table 4, we see the same effect. In addition, we see that the larger the subgroup, the easier it is for φ_{WRL} to distinguish the exceptional sequences from the other sequences. For instance, for a subgroup where $order = 4$, and when $Z = 20$, the median rank increases from 1 to 2, to 3, and finally to 18 when the subgroup size decreases from 50% to 35%, to 25%, and to 13%.

Results for φ_{WD} and ω_{tv} can be found in the repository.[3] Essentially, we find that ω_{tv} is fairly robust for subgroup size. For φ_{WD}, we see a pattern: the smaller the Z,

the easier it is to find the true subgroup. The subgroup size does not seem to influence the ranking.

In sum, our quality measures based on information-theoretic scoring functions give stable results for ground truth subgroups of varying size. For very large subgroups that contain more than 50% of the sequences, the rank increases slightly but not worryingly much.

6 Experiments on real-world data

6.1 Continuous glucose measurements in DIALECT-2

We now further analyse sequential data from the second DIAbetes and LifEstyle Cohort Twente (DIALECT-2) (Gant et al. 2017; Den Braber et al. 2021), already shortly introduced in Sect. 2. Our interest is in analysing measurements from the FreeStyle Libre sensor, an intermittently continuous glucose monitoring (iCGM) sensor. Monitoring blood glucose values using an iCGM device may become the new way to monitor glycemic control for patients with diabetes type 2 (Danne et al. 2017; Den Braber et al. 2021). The current clinical accepted standard is monitoring of glycated haemoglobin (HbA1c), a measure that is linearly related to the average blood glucose concentration of the past few months (World Health Organization et al. 2011) and known to increase the risk of comorbidities. However, monitoring HbA1c does not help to reduce hypoglycemic episodes and it does not reflect blood glucose fluctuations well enough (Kovatchev et al. 2003).

In general, blood glucose values are considered to be in the desired range (IR) if they are between 3.9 and 10.0 mmol/L. Danne et al. (2017) furthermore distinguish blood glucose values that are *below* (BR) and *above* range (AR). These lower and upper ranges are again subdivided into BR_1 (3.0–3.9 mmol/L), BR_2 (<3.0 mmol/L), AR_1 (10.0 - 13.9 mmol/L) and AR_2 (>13.9 mmol/L).

The DIALECT-2 dataset contains the information of 126 patients, with an average sequence length of $T = 1210$ (SD: 158). Not all sequences have the same length because sometimes patients forget to upload the stored data or to charge the iCGM-device. On average, 55 (SD: 38) measurements were missing, and no patient had more than 312 missing values. As numerical descriptive attributes, we use age, diabetes duration, body mass index, waist/hip ratio, predicted muscle mass, systolic blood pressure, diastolic blood pressure, heart rate, alcohol intake and smoking pack years. The binary descriptors are sex, whether or not someone uses insulin, and if so, with what type of scheme, whether or not someone uses metformin, repaglinide or sulphonylurea, the presence of micro vascular disease and the presence of macro vascular disease. We use one ordinal descriptive attribute: HbA1c category. A HbA1c value $\leq$ 53 mmol/mol is considered *low*, a value from 54 to 62 mmol/mol *medium* and a value $\geq$ 63 mmol/mol *high* (e.g. McGuire et al. 2016; Battelino et al. 2019).

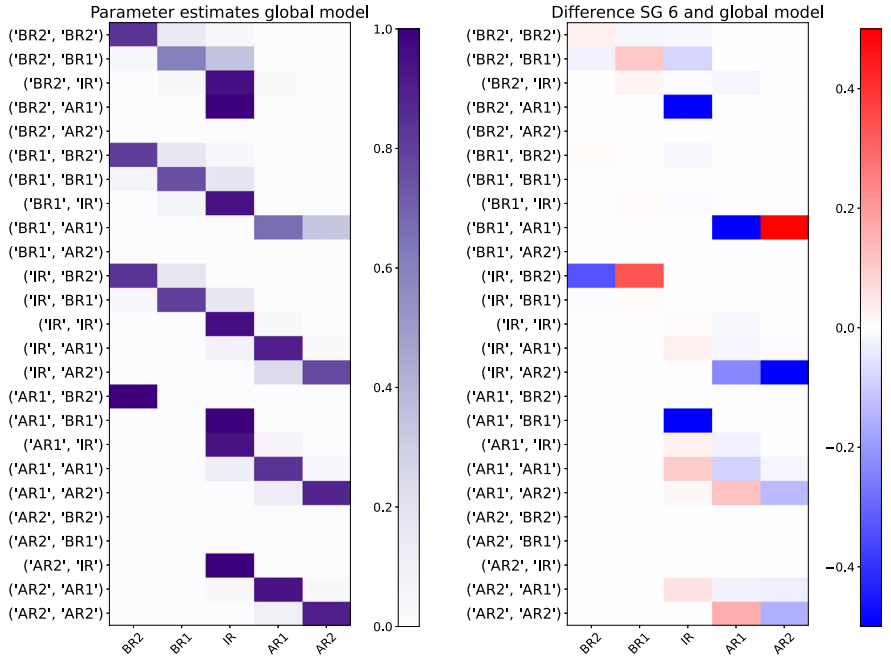

Fig. 6 Parameter estimates of the global model (left; reproduction of Fig. 1a for purposes of easy comparison with the figure on the right) and the difference between the sixth best-scoring subgroup and the global model (right). Description is *HbA1c category* = *low* $\wedge$ *alcohol intake* $\leq$ *25 units/month*. Coverage: 24%

6.1.1 Long sequences of discrete blood glucose level

We analyse the sequences of blood glucose values in two ways. First, we discretise the continuous glucose measurements into the 5 blood glucose levels as discussed before. As a quality measure, we use φ_{AIC} since our findings from controlled experiments indicate that this measure is more robust for the number of observations than φ_{BIC} (Sect. 5) and performs similarly to φ_{AICc}.

We apply the extended beam search algorithms (Sect. 4.2) with parameters $w = 25$, $d = 3$ and $q = 20$. Descriptive attributes are refined with standard strategies (cf. Duivesteijn et al. (2016)), where we treat numerical attributes with the dynamic discretisation strategy lbca[4] from Meeng and Knobbe (2021) using $b = 4$ bins. Furthermore, we set a minimum subgroup size of 10% and use start parameter $s = 4$.

It is generally known that the beam search algorithm can discover redundant subgroups. Therefore, we implement three techniques as proposed by Van Leeuwen and Knobbe (2012). First, we perform description-based selection with a fixed-size of $2w = 50$. Here, subgroups are skipped if they have 1) equal quality and 2) the same description except for 1 condition. Second, we perform fixed-size cover-based beam selection where the quality value of a subgroup is weighted based on how many

[4] Here, lbca is a concatenation of *Local* discretisation timing, *Binary* interval type, *Coarse* granularity, and *All* selection method.

instances (i.e. sequences) were already covered by another subgroup (Lavrač et al. 2004; Van Leeuwen and Knobbe 2012). The weight of subgroup SG is defined as

$$w^{SG} = \frac{1}{N^{SG}} \sum_{r=1}^{N^{SG}} \gamma^{c_r}$$

where count c_r is the number of times that sequence r is covered by other subgroups. We set $\gamma = 0.9$. Third, we apply dominance pruning to the result list.

We find that the entire dataset is best modelled using a 2nd order Markov chain (Alg. 1), where we see both diagonal patterns and unusual fluctuations such as blood glucose values changing from IR $\rightarrow$ AR_2 and from AR_1 $\rightarrow$ BR_2 (left plot in Figure 6).

The top-20 subgroups are best fitted with either a 1st or 2nd order Markov chain. The first subgroup contains 18% of the patients with description *HbA1c category* = *low* $\wedge$ *diabetes duration* $\leq$ *20 years* $\wedge$ *21.3* $\leq$ *BMI* $\leq$ *35.7*. The subgroup's parameter estimates were already shown in Fig. 1b. Actually, the conditions for diabetes duration and BMI cover all patients that are in the first three quartiles of the respective variable distributions; they only remove a few extreme patients from the full subgroup. However, the first condition that selects patients with a low HbA1c value is very interesting as we know that HbA1c correlates with the average blood glucose concentration of the past few months and increases the risk for comorbidities (World Health Organization et al. 2011).

For the first subgroup, we find a strong diagonal transition pattern (i.e. people tend to stay at the same blood glucose level) and the blood glucose values of these patients fluctuate less than those in the overall patient population. This may also be the reason that a 1st order Markov chain suffices. The second-best-scoring subgroup selects patients with a *high* HbA1c value. Figure 7 shows the difference in parameter estimates between subgroup 2 and subgroup 1. It is immediately visible that patients in subgroup 2 are more likely to transition towards higher blood glucose levels than patients in subgroup 1 (see red squares for BR_1 $\rightarrow$ IR and IR $\rightarrow$ AR_1) and less likely to transition to lower levels (see blue squares). Since HbA1c is known to correlate with average blood glucose values, these findings are confirmed by clinicians and domain experts.

The right plot in Fig. 6 presents the difference in parameter estimates between the sixth subgroup and the global model. Both were best fitted with a 2nd order Markov chain. Here, like for subgroup 1, patients with *low* HbA1c values are selected. Although we see more fluctuations than for subgroup 1, we see a similar trend where blood glucose levels are likely to either stay the same (see for instance the red squares in the first and last two rows), or transition towards a lower blood glucose level (see the red squares for (AR_1, AR_1) $\rightarrow$ IR and (AR_1, AR_2) $\rightarrow$ AR_1).

6.1.2 Short sequences of TIR, TBR, and TAR

For our second analysis, we derive the percentage per day that a patient has blood glucose values at level BR_2, BR_1, IR, AR_1, or AR_2. This is referred to as the Time

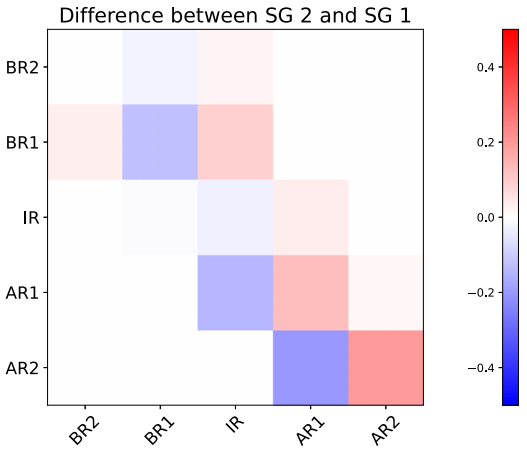

Fig. 7 Difference between parameter estimates of second best-scoring subgroup and first best-scoring subgroup (Fig. 1b). Description is *HbA1c category = high* $\wedge$ *30.9* $\leq$ *fat percentage* $\leq$ *60.3* $\wedge$ *118.7* $\leq$ *syst.bp* $\leq$ *158.7*. Coverage: 24%

In Range (TIR), Time Below Range (TBR), and Time Above Range (TAR) (Danne et al. 2017; Battelino et al. 2019; Den Braber et al. 2021). For each of these values, we compare the percentages with the guidelines (see Table 5) (Danne et al. 2017). Subsequently, for each day, we assign one out of 8 state-values (see Table 5). This gives us one sequence of length $T = 14$ per patient.

The entire dataset is best modelled with a 1st order Markov chain, because the sequences are relatively short and the dataset size is relatively small. In general, patients stay in or move towards state AC (low TIR, good TBR, high TAR) or state AH (good TIR, good TBR, good TAR) (top left plot in Fig. 8).

The amount of available data for the subgroups is even smaller than for the entire dataset, and it is therefore not possible to find subgroups with higher order Markov chains. The first subgroup contains 24% of the patients with description *18.2* $\leq$ *fat percentage* $\leq$ *42.2* $\wedge$ *34.6 kg* $\leq$ *predicted mean mass* $\leq$ *65.5 kg* $\wedge$ *HbA1c category = high*. These patients are likely to transition to state AA, AC and AG (see top right plot in Fig. 8, red columns), which corresponds to the situation where TAR is too high.

The third best-scoring subgroups covers patients with, among others, *HbA1c = low* (bottom left plot in Fig. 8). Here, we see transitions AC $\rightarrow$ AG (TIR is good instead of low) and AE $\rightarrow$ AF (TAR is good instead of high).

The fourteenth best-scoring subgroup covers patients with a high HbA1c value, who are additionally older than the average patient. These patients not only have a TAR that is too high, but they are also less likely to have a TIR that is good. We see this in the bottom right plot in Fig. 8 by the blue columns, and by the two red columns for state AA and state AC. Clinicians and domain experts confirm these findings. It is generally accepted that the blood glucose values of older patients are a bit higher since their risk for comorbidities is lower and their life expectancy shorter.

Table 5 Conversion of time spent in glucose level ranges into states suitable for Markov chains

(a) Time In Range (TIR), Time Below Range (TBR) and Time Above Range (TAR) are calculated based on whether glucose values are IR, BR_1, BR_2, AR_1 or AR_2.

TIR	IR $< 70\%$	IR $\geq 70\%$
	low	good
TBR	$BR_1 < 4\%$	$BR_1 \geq 4\%$
$BR_2 < 1\%$	good	high
$BR_2 \geq 1\%$	high	high
TAR	$AR_1 < 25\%$	$AR_1 \geq 25\%$
$AR_2 < 5\%$	good	high
$AR_2 \geq 5\%$	high	high

(b) Eight state-values are created based on the combination of the TIR, TBR and TAR.

State	TIR	TBR	TAR
AA	low	high	high
AB	low	high	good
AC	low	good	high
AD	low	good	good
AE	good	high	high
AF	good	high	good
AG	good	good	high
AH	good	good	good

Medically inspired cut-off percentages are taken from Danne et al. (2017)

6.2 MovieLens

Finally, we analyse the MovieLens 100K dataset.[5] The dataset consists of 943 users, each rating at least 20 movies on an integer scale from 1 to 5. We consider sequences of ratings per user where $20 \leq T \leq 737$ with an average (SD) sequence length of $T = 203$ (139). The Markov chain uses the movie rating values as its state space, so we have $V = \{1, 2, 3, 4, 5\}$ and $S = 5$. Specifically, we search for subgroups of users with exceptional rating patterns based on demographic information (age, gender, occupation) and sequence information (sequence length T). The idea behind using sequence length as a descriptive attribute is to form subgroups of users who rate a lot, or subgroups of users who rate relatively little. As described earlier in Sect. 2.2, a user's rating sequence is either entirely part of a subgroup, or not; we do not split sequences.

The extended beam search algorithms are performed with parameters $q = 20$, $w = 25$, $d = 3$, $b = 4$, and $s = 4$. Subgroups should cover at least 10% of all users. We adopt the same redundancy strategies as in Sect. 6.1 with $\gamma = 0.9$ and the

[5] MovieLens 100K dataset is available at https://grouplens.org/datasets/movielens/.

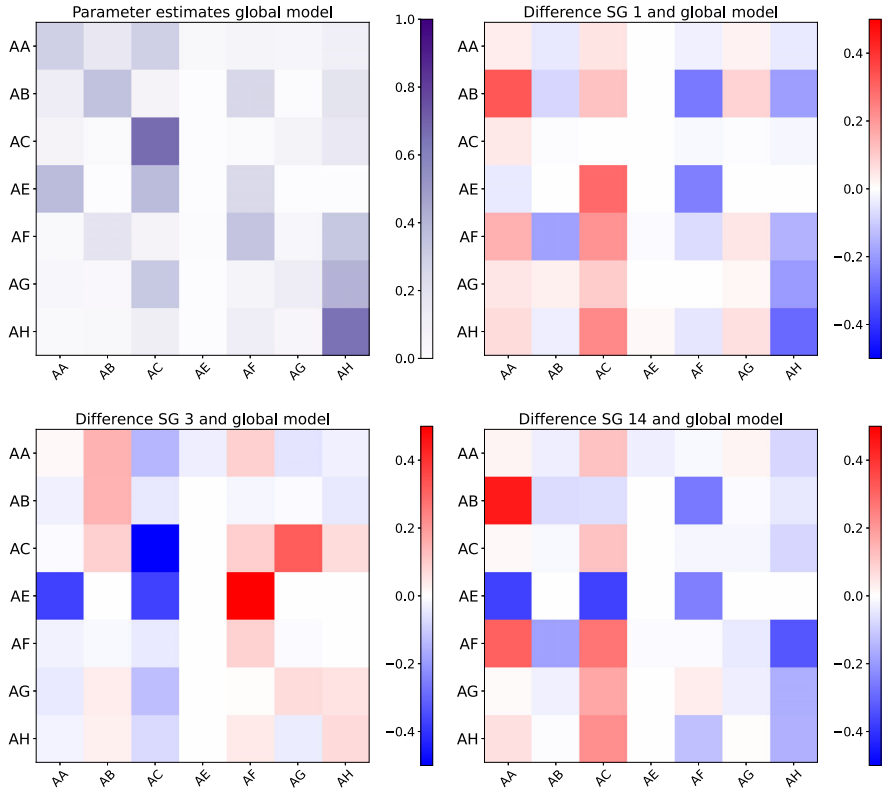

Fig. 8 Parameter estimates of the global model (top left) and the difference between three subgroups and the global model. Top right: First best-scoring subgroup with description *HbA1c category = high* $\wedge$ *18.2* $\leq$ *fat percentage* $\leq$ *42.2* $\wedge$ *34.6 kg* $\leq$ *predicted mean mass* $\leq$ *65.5 kg*. Coverage: 24%. Bottom left: Third best-scoring subgroup with description *HbA1c category = low* $\wedge$ *alcohol intake* $\leq$ *18 units/month*. Coverage: 22%. Bottom right: Fourteenth best-scoring subgroup with description *HbA1c category = high* $\wedge$ *67* $\leq$ *age* $\leq$ *84* $\wedge$ *0.9* $\leq$ *waist/hip ratio* $\leq$ 1.2. Coverage: 21%

description-based selection procedure with a fixed-size of $2w = 50$. Again, we use quality measure φ_{AIC} to evaluate candidate subgroups.

The entire dataset is best fitted with a 2$^{\text{nd}}$ order Markov chain. In the top-20, we find subgroups of users with either a 1$^{\text{st}}$ or a 3$^{\text{rd}}$ order Markov chain. For instance, the best-scoring subgroup selects users with *occupation* $\neq$ {other, technician} $\wedge$ *183* $\leq$ *sequence length* $\leq$ *737* (cov: 16%) and is best fitted with a 3$^{\text{rd}}$ order Markov chain. In subgroup 2, on the other hand, users with short sequences where *20* $\leq$ *sequence length* $\leq$ *73* $\wedge$ *occupation* $\neq$ *technician* are selected (cov: 52%) and a 1$^{\text{st}}$ order Markov chain is fitted.

The results show that a quality measure that takes into account the number of free parameters, such as φ_{AIC}, allows for a flexible evaluation of candidate subgroups. It is reasonable to assume that the entire MovieLens dataset is fitted with a 2$^{\text{nd}}$ order Markov chain as a compromise between shorter and longer sequences. Evaluating candidate subgroups based on such a 2$^{\text{nd}}$ order model (as would be done in the traditional EMM

framework using a parameter-based quality measure) would reduce the probability of finding patterns in subgroups that contain short, or long, sequences. Although the MovieLens dataset seemingly does not encompass meaningful relations between user demographics and sequence length, such patterns may exist in other datasets and can be searched for using quality measures based on information-theoretic scoring functions.

7 Discussion

We proposed a method for mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. On average, the quality measures based on information-theoretic scores outperform the other measures; they give a higher rank to ground truth subgroups, they find the correct Markov chain order more often and they are able to detect subgroups that would otherwise not have been found (Sect. 5.1). In datasets with many, short sequences, exceptional starting behaviour can be detected (Sect. 5.2). For long sequences, our quality measures perform robustly w.r.t. the order of the global model, the start parameter, the subgroup size, and the description length (Sect. 5.3).

In some situations, other quality measures can be valuable as well. For instance, if subgroups are expected to have a similar Markov chain order as the global model, quality measure ω_{tv} performs fine (but the information-theoretic based measures do not perform worse). In situations where the subgroups are expected to have a (much) higher Markov chain order than the global model, semi-evaluation measure φ_{WRL} can be used. Note that φ_{WRL} requires a relatively large number of observations in order to extract the subgroup. The performance of quality measure φ_{WD} is a bit unpredictable, possibly due to its sensitivity to start parameter s.

The quality measures based on information-theoretic scoring functions are flexible and can detect subgroups whose Markov chains 1) have the same order as the global model and 2) have a deviating order. In practice, the global model is an average over all sequences in a dataset, and quality measures that use a penalty based on the number of observations M and the number of free parameters K are able to go beyond such an average. In our study, we have chosen three common penalties; AIC, AIC with small sample correction, and BIC. However, our proposed EMM framework allows for the extension to other penalised scoring functions in a straightforward way.

Interestingly, our findings from controlled experiments do not show much difference between φ_{AIC}, φ_{AICc}, and φ_{BIC}. The first two slightly outperform φ_{BIC} when the number of observations and the number of free parameters is large (Sect. 5.1), due to the excessive penalisation by the BIC scoring functions.

It is a bit unexpected that we do not see a difference between φ_{AIC} and its variant for small sample sizes φ_{AICc}. When $1 < \frac{M}{K} < 40$, the penalty in φ_{AICc} is supposed to do more justice to the uncertainty of parameter estimates than the penalty in φ_{AIC} (Sugiura 1978; Hurvich and Tsai 1995; Burnham and Anderson 2004). This means that we would expect φ_{AICc} to give a larger penalty than φ_{AIC}. A possible explanation for the absence of the effect of such a penalty is that as soon as the true subgroup is found, it is so distinctive from the other sequences that a larger penalty does not bother the ranking. Another possible reason could be that with our simulation parameters, we

have not been able to capture the dataset characteristics for which such a penalty would make a difference. Nevertheless, in our synthetic data experiments there are many subgroups where $\frac{M}{K} > 40$ and then, the difference between AIC and AICc disappears nonetheless (Sugiura 1978; Hurvich and Tsai 1995; Burnham and Anderson 2004).

For all our experiments, we use the extended beam search algorithm as presented in Sect. 4.2. It is generally known that beam search may discover redundant subgroups. Therefore, we apply three methods from Van Leeuwen and Knobbe (2012) during our real-world data experiments (see Sect. 6.1). Note that in the synthetic data experiments, we designed the simulation such that the descriptive attributes do not overlap in coverage (i.e. binary only). Hence, redundancy did not play a role and we were able investigate the ranking of the ground truth subgroup in a more controlled manner.

We performed the description-based selection using a fixed size of $2w$. In our implementation, description-based selection of candidate subgroups occurs before cover-based selection. We think that it is reasonable to assume that starting the latter with $2w$ subgroups allows for a beam that contains w diverse subgroups. We furthermore decided that γ should not be too small in order to not be too rigorous with decreasing the quality of subgroups that have redundant coverage. We therefore set $\gamma = 0.9$.

The beam search algorithm requires a set of parameters that may come across as arbitrary. In general, we suggest to choose the parameter values such that the result list is practical and meaningful. For one thing, this means that the result list should be diverse (Van Leeuwen and Knobbe 2012), but it also means that subgroup descriptions should not be too long or a subgroup should not be too small. We have chosen to set $d = 3$ in order to allow for descriptions that contain at most three attributes. These descriptions can easily be remembered and interpreted by the domain expert and are therefore practical. Furthermore, subgroups should be substantially large in order to adopt separate policies or treatment schemes; it seems reasonable to form subgroups that cover at least 10% of the population. Because in both synthetic and real-world data experiments, the number of descriptive attributes is relatively small, we deemed $w = 25$ to allow for sufficient exploration of the search space. For much higher dimensional datasets, possibly this parameter can be increased at some additional computational expense. Last, parameter q is often determined in consultation with domain experts. In our experience, a top-20 result list is not too long to prevent interpretation but long enough to find valuable subgroups. Note that changing q will not actually change the results; it merely specifies the cutoff point in a list of ordered subgroups.

8 Conclusion

We proposed a method for mining sequences with exceptional transition behaviour of varying order. Specifically, we use the framework of Exceptional Model Mining (EMM) to find subgroups of sequences and propose a model class for varying order Markov chains. Our model class allows for discovering subgroups in situations where the order of the Markov chain differs between the subgroup and the dataset. Such a situation requires the comparison of a different number of parameters. We therefore do not use a parameter-based quality measure as is common in EMM, but propose three

new quality measures based on information-theoretic scoring functions: φ_{AIC}, φ_{BIC}, and φ_{AICc}.

Our findings from controlled experiments show that all three quality measures find exceptional transition behaviour of varying order. They all give a first rank to the ground truth subgroup when sequences have a length $T \geq 50$. For shorter sequences, the ability to give a first rank to the ground truth subgroup depends on the state-space, the descriptive space and the ground truth Markov chain order. Naturally, the higher the Markov chain order of the subgroup, the more observations are needed. Nevertheless, φ_{AIC}, φ_{BIC}, and φ_{AICc} all seem sensitive enough to detect the correct Markov chain order but sensitive enough to prevent overfitting. Compared to φ_{BIC}, we find that φ_{AIC} is slightly more robust for the number of observations. We have not seen important differences between φ_{AIC} and φ_{AICc}.

We furthermore add to existing work by seeking for subgroups of sequences, as opposite to subgroups of transitions. In particular, we say that sequence-level descriptive attributes contain information about entire sequences. As such, we are able to form subgroups of homogeneous sequences that are heterogeneous with respect to the rest of the dataset. In contrast, transition (or time) level descriptors only partly include a sequence in the subgroup, which is not meaningful in situations where we want to identify the originators of sequences, such as patients or user-sessions.

The practical relevance of our approach is shown using data from an observational study of adult persons with diabetes type 2 (Sect. 6.1). In the first experiment, a 2nd order Markov chain is fitted to the entire dataset and we find subgroups of either a 1st or 2nd order Markov chain. For instance, we find a first subgroup that covers patients with low HbA1c values, a measure known to correlate with average blood glucose values. The subgroup is best modelled with a 1st order Markov chain and its parameter estimates show an increased probability of staying in or moving towards desired blood glucose values. Clinicians and domain experts confirmed that the blood glucose values of these type of patients fluctuate less.

In the second experiment, we find, among others, subgroups covering patients with high HbA1c values and an above average age. The model parameters indicate an increased probability of transitioning to blood glucose values that are too high. Clinicians and domain experts confirmed these findings, and furthermore add that it is generally accepted that the blood glucose values of older patients are a bit higher since their risk for comorbidities is lower and their life-expectancy shorter.

Acknowledgements This research is supported by EDIC project funded by NWO. We thank the EDIC consortium and the ZGT hospital for allowing us to analyse the data from the DIALECT-2 study. We especially thank Niala Den Braber (PhD candidate at Universiteit Twente and researcher internal medicine at ZGT hospital) and prof. dr. Goos Laverman (internist-nephrologist at ZGT hospital) for giving us clinical valuation of our findings. In addition, we thank our colleagues dr. Robert Peharz for giving us useful insights on Markov chains and DBNs and dr. Maryam Tavakol for guiding us towards the MovieLens dataset.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Akaike H (1973) Information theory and the maximum likelihood principle. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT), pp. 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control (TACON) 19(6):716–723

Battelino T, Danne T, Bergenstal RM, Amiel SA, Beck R, Biester T, Bosi E, Buckingham BA, Cefalu WT, Close KL et al (2019) Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. Diabetes Care (DC) 42(8):1593–1603

Becker M, Lemmerich F, Singer P, Strohmaier M, Hotho A (2017) MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data. Data Min Knowl Discov (DAMI) 31(5):1359–1390

Bosc G, Boulicaut JF, Raïssi C, Kaytoue M (2018) Anytime discovery of a diverse set of patterns with Monte Carlo Tree Search. Data Min Knowl Discov (DAMI) 32(3):604–650

Bueno MLP, Hommersom A, Lucas PJ, Janzing J (2019) A probabilistic framework for predicting disease dynamics: a case study of psychotic depression. J Biomed Inf (JBI) 95:103232

Bueno MLP, Hommersom A, Lucas PJ (2020) Temporal exceptional model mining using dynamic Bayesian networks. In: International Workshop on Advanced Analytics and Learning on Temporal Data (AALTD), Springer, pp. 97–112

Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res (SMR) 33(2):261–304

Dagum P, Galper A, Horvitz E (1992) Dynamic network models for forecasting. In: Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI), Elsevier, pp. 41–48

Danne T, Nimri R, Battelino T, Bergenstal RM, Close KL, DeVries JH, Garg S, Heinemann L, Hirsch I, Amiel SA et al (2017) International consensus on use of continuous glucose monitoring. Diabetes Care (DC) 40(12):1631–1640

Den Braber N, Vollenborek-Hutten MMR, Westerik KM, Bakker SJL, Navis G, van Beijnum BJF, Laverman GD (2021) Glucose regulation beyond HbA1c in type 2 diabetes treated with insulin: Real-world evidence from the DIALECT-2 cohort. Diabetes Care (DC) 44:2238–2244

Deng J, Kang B, Lijffijt J, Bie TD (2020) Explainable subgraphs with surprising densities: A Subgroup Discovery approach. In: Proceedings of the SIAM International Conference on Data Mining (SDM), pp. 586–594

Duivesteijn W, Feelders A, Knobbe A (2012) Different slopes for different folks: Mining for exceptional regression models with Cook's distance. In: Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining (KDD), pp. 868–876

Duivesteijn W, Feelders AJ, Knobbe A (2016) Exceptional Model Mining. Data Min Knowl Discov (DAMI) 30(1):47–98

Gant CM, Binnenmars SH, Berg EVd, Bakker SJ, Navis G, Laverman GD (2017) Integrated assessment of pharmacological and nutritional cardiovascular risk management: blood pressure control in the DIAbetes and LifEstyle Cohort Twente (DIALECT). Nutrients 9(7):709

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc (JASA) 102(477):359–378

Herrera F, Carmona CJ, González P, Del Jesus MJ (2011) An overview on subgroup discovery: Foundations and applications. Knowl Inf Syst (KAIS) 29(3):495–525

Hurvich CM, Tsai CL (1995) Model selection for extended quasi-likelihood models in small samples. Biometrics 55:1077–1084

Jaroszewicz S (2010) Using interesting sequences to interactively build Hidden Markov Models. Data Min Knowl Discov (DAMI) 21(1):186–220

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc (JASA) 90(430):773–795
Kiseleva J, Lam HT, Pechenizkiy M, Calders T (2013) Predicting current user intent with contextual Markov models. In: IEEE international conference on data mining workshops. IEEE, pp 391–398
Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Proceedings of the Conference on Advances in Knowledge Discovery and Data Mining (PAKDD). AAAI/MIT Press, pp. 249–271
Kovatchev BP, Cox DJ, Kumar A, Gonder-Frederick L, Clarke WL (2003) Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data. Diabetes Technol Ther (DTT) 5(5):817–828
Lavrač N, Kavšek B, Flach P, Todorovski L (2004) Subgroup discovery with CN2-SD. J Mach Learn Res 5(Feb):153–188
Leman D, Feelders A, Knobbe A (2008) Exceptional model mining. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD). Springer, pp. 1–16
Lemmerich F, Becker M, Atzmueller M (2012) Generic pattern trees for exhaustive Exceptional Model Mining. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD). Springer, pp. 277–292
Lemmerich F, Becker M, Singer P, Helic D, Hotho A, Strohmaier M (2016) Mining subgroups with exceptional transition behavior. In: Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining (KDD), pp. 965–974
Lijffijt J, Kang B, Duivesteijn W, Puolamaki K, Oikarinen E, De Bie T (2018) Subjectively interesting subgroup discovery on real-valued targets. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 1352–1355
Mathonat R, Nurbakova D, Boulicaut JF, Kaytoue M (2021) Anytime mining of sequential discriminative patterns in labeled sequences. Knowl Inf Syst (KAIS) 63(2):439–476
McGuire H, Longson D, Adler A, Farmer A, Lewin I (2016) Management of type 2 diabetes in adults: Summary of updated NICE guidance. BMJ, 353
Meeng M, Knobbe AJ (2021) For real: a thorough look at numeric attributes in subgroup discovery. Data Min Knowl Discov 35(1):158–212
Meier J, Dietz A, Boehm A, Neumuth T (2015) Predicting treatment process steps from events. J Biomed Inf (JBI) 53:308–319
Mollenhauer D, Atzmueller M (2020) Sequential exceptional pattern discovery using pattern-growth: an extensible framework for interpretable machine learning on sequential data. In: Proceedings of the International Workshop on Explainable and Interpretable Machine Learning (XI-ML)
Peharz R, Kapeller G, Mowlaee P, Pernkopf F (2014) Modeling speech with sum-product networks: application to bandwidth extension. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 3699–3703
Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu MC (2004) Mining sequential patterns by pattern-growth: the PrefixSpan approach. IEEE Trans Knowl Data Eng (TKDE) 16(11):1424–1440
Pirolli PL, Pitkow JE (1999) Distributions of surfers' paths through the world wide web: empirical characterizations. World Wide Web 2(1–2):29–45
Pohle J, Langrock R, van Beest FM, Schmidt NM (2017) Selecting the number of states in Hidden Markov Models: pragmatic solutions illustrated using animal movement. J Agric Biol Environ Stat (JABES) 22(3):270–293
Sadagopan N, Li J (2008) Characterizing typical and atypical user sessions in clickstreams. In: Proceedings of the international conference on World Wide Web (WWW), pp. 885–894
Sarukkai RR (2000) Link prediction and path analysis using Markov chains. Comput Netw 33(1–6):377–386
Schoof J, Pryor S (2008) On the proper order of Markov chain model for daily precipitation occurrence in the contiguous united states. J Appl Meteorol Climatol (JAMC) 47(9):2477–2486
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Singer P, Helic D, Taraghi B, Strohmaier M (2014) Detecting memory and structure in human navigation patterns using Markov chain models of varying order. PloS one 9(7):e102070
Song H (2017) Model-based subgroup discovery. PhD thesis, University of Bristol
Song H, Flach P, Kalogridis G (2015) Dataset shift detection with model-based subgroup discovery. In: International Workshop on Learning over Multiple Contexts (LMCE)
Song H, Kull M, Flach P, Kalogridis G (2016) Subgroup discovery with proper scoring rules. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD). Springer, pp. 492–510

Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. Commun Stat Theory Methods 7(1):13–26
Tong H (1975) Determination of the order of a Markov chain by Akaike's information criterion. J Appl Probab 12(3):488–497
Van Leeuwen M, Knobbe A (2012) Diverse subgroup set discovery. Data Min Knowl Discov 25(2):208–242
Wilks DS (1999) Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. Agric For Meteorol 93(3):153–169
World Health Organization, et al. (2011) Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. No. WHO/NMH/CHP/CPM/11.1, World Health Organization
Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD). Springer, pp 78–87
Zucchini W, MacDonald IL, Langrock R (2017) Hidden Markov models for time series: an introduction using R. CRC Press, Boca Raton