Staged trees for discrete longitudinal data

Jack Storror Carter[1*], Manuele Leonelli[2], Eva Riccomagno[1], Alessandro Ugolini[3]

[1]Dipartimento di Matematica, Università degli Studi di Genova, Genova, Italy.
[2]School of Science and Technology, IE University, Madrid, Spain.
[3]Department of Surgical and Diagnostic Sciences, University of Genoa, Genoa, Italy.

*Corresponding author(s). E-mail(s): jack.carter@dima.unige.it;
Contributing authors: manuele.leonelli@ie.edu;
riccomagno@dima.unige.it; alessandro.ugolini@unige.it;

Abstract

In this paper we investigate the use of staged tree models for discrete longitudinal data. Staged trees are a type of probabilistic graphical model for finite sample space processes. They are a natural fit for longitudinal data because a temporal ordering is often implicitly assumed and standard methods can be used for model selection and probability estimation. However, model selection methods perform poorly when the sample size is small relative to the size of the graph and model interpretation is tricky with larger graphs. This is exacerbated by longitudinal data which is characterised by repeated observations. To address these issues we propose two approaches: the longitudinal staged tree with Markov assumptions which makes some initial conditional independence assumptions represented by a directed acyclic graph and marginal longitudinal staged trees which model certain margins of the data.

Keywords: Chain event graphs; Discrete data; Longitudinal studies; Staged trees

1 Introduction

Longitudinal studies are characterised by the repeated observation of individuals over time. This differs to cross-sectional studies in which each individual is only

observed at a single time point. Longitudinal studies benefit from increased power over cross-sectional studies (Zeger and Liang, 1992) and are able to answer more sophisticated questions because the data shows how individuals change over time. However, they require additional modeling considerations due to the dependence between the repeated observations.

Longitudinal studies are often viewed in a regression context in which one variable is considered the outcome and the remaining variables the covariates. The aim is then to determine the effect of the covariates on the outcome. We denote the repeated observations of the *longitudinal outcome* by Y_t, $t = 1 \ldots, T$ and of the *longitudinal covariates* by X_t, $t = 1 \ldots, T$ where T is the total number of time points. The study may also include covariates whose value does not change over time. We call these *time invariant covariates* and denote them by Z. In this paper we assume that all variables are discrete with a finite sample space. The finite sample space may be either categorical or ordinal.

Traditional methods for such longitudinal data typically involve a generalised linear model with three approaches being particularly popular - the marginal model, transition model and random effects model. Each of these approaches offer different methods for modeling both the effect of the covariates on the outcome and the dependence between the outcome at different time points. Further details of these models will be given in Section 4.

In this paper, we instead investigate the use of staged tree models for discrete longitudinal data. Staged trees, first introduced by Smith and Anderson (2008) (see Collazo et al. (2018) for a detailed review) are a type of probabilistic graphical model represented by a coloured directed tree graph. They also have a more compact graphical representation called the chain event graph, obtained through a coalescence of the vertices. The main feature of staged tree models is the ability to identify events that have identical probability distributions. For example, in a regression setting they can specify different values of the covariates for which the probability distribution of the outcome is the same. Additionally, a temporal ordering of variables is often implicitly assumed in staged tree models and so longitudinal data easily fits into the staged tree framework.

While including a temporal element has been explored in staged trees - for example, the dynamic chain event graph and the inclusion of conditional holding times (Freeman and Smith, 2011a; Barclay et al., 2015) - and longitudinal data has been modelled using staged trees in applications (Hutton, 2015), no specific methods for using staged trees with the form of longitudinal data described above have been proposed. Staged trees are often compared to Bayesian networks - another type of probabilistic graphical model - for which some methods for longitudinal data have been proposed. For example, the dynamic Bayesian network for longitudinal data (McGeachie et al., 2016; Prinzie and Van den Poel, 2011; Chen et al., 2012) and improved probability estimation for longitudinal data (Bellazzi and Riva, 1998).

Regression based methods make efficient use of the data by making strong assumptions - namely that the effect of the covariates on the outcome follows a specific linear model and the dependence between outcomes at different times follows a specific form. The inferences obtained often rely on these assumptions which can at times be hard to

interpret or test. When selected by data, staged tree models do not require any such initial assumptions. However, this comes with the drawback that they require a large sample size relative to the number of variables and the size of their sample spaces. This is critical with longitudinal data because observing an additional time point results in an increase in the number of variables in the model without an increase in the sample size.

While existing staged tree methodology can be applied to longitudinal data, in this paper we suggest two additional approaches for smaller sample sizes. One makes some initial conditional independence assumptions which can be represented via a directed acyclic graph. These assumptions are easy to interpret due to their graphical representation and could also be tested using data via standard methods for Bayesian networks (see e.g. Scutari and Denis, 2021). The other approach models certain marginal distributions with staged trees. While this no longer allows estimation of the full joint probability distribution, one is still able to determine how the effect of the covariates on the outcome changes over time.

The remainder of the paper is organised as follows. In Section 2 we introduce staged tree models and give a brief summary of current model selection methods for staged trees. Section 3 introduces three proposed approaches for using staged trees to model longitudinal data - the full longitudinal staged tree, the longitudinal staged tree with Markov assumptions and marginal longitudinal staged trees. There is also discussion about the choice of prior parameters and the robustness of these methods to small sample sizes. In Section 4 we compare staged tree models with traditional regression methods and discuss some of the benefits of using staged tree models for longitudinal data. We finish in Section 5 by applying the proposed staged tree methods to a real data set about tooth decay in children.

2 Staged trees and chain event graphs

2.1 Definition

In this section we present staged tree models and their compact representation chain event graphs with the help of an example. This example will later be extended to demonstrate our proposed methods for longitudinal data. The example first appeared in Koch et al. (1977) and it compares a new drug for treating depression with a standard treatment. Individuals were randomly assigned one of the treatments and were also diagnosed to have either mild or severe depression. After one week of treatment the individual's symptoms were assessed to be either normal (N) or abnormal (A).

A *staged tree* (ST) is a probability tree with an equivalence relation on its non-leaf vertices. More specifically, a ST is a probabilistic graphical model for a process consisting of a sequence of discrete events. To construct a ST we begin with an *event tree* containing a single *root* vertex with no incoming edges and at least two outgoing edges, representing the start of the process, and a number of *leaf* vertices which have a single incoming edge and no outgoing edges, representing the end of the process. An event tree for the depression data is in Figure 1a where the root vertex is labelled s_0 and the leaf vertices are $s_7 - s_{14}$. All remaining vertices in the graph have a single incoming edge and at least two outgoing edges.

All non-leaf vertices (including the root) are called *situations* and represent a possible state at which the process can arrive. The edges in the event tree are labeled such that for each situation, the outgoing edge labels describe all possible events that can occur at the next stage of the process. As such, the event tree fully describes the sample space of the process. For example, the vertex s_2 in Figure 1a represents the situation in which an individual has been assigned to the standard treatment. The subsequent event is the diagnosis which can be either Mild or Severe. Edges in Figure 1a are additionally labelled with the number of observations of each event.

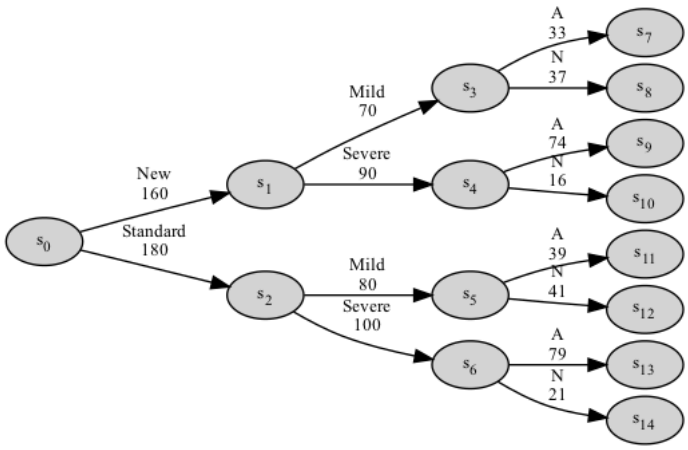
For each situation in the graph, one can associate a probability distribution, called the transition probabilities, representing the conditional probabilities of the subsequent event of the process. One obtains a joint distribution for the whole process by assigning a probability distribution to all situations and then using the standard chain rule of probability.

The most general statistical model places no further constraints on the probability distributions at each situation. However, a ST model restricts the space by assuming that some situations (with the same outgoing edge labels) have the same probability distribution. When this is the case, the two situations are said to be in the same *stage*. This is represented graphically by colouring vertices according to which stage they are in (with any singleton stages coloured white for simplicity). For example, in Figure 1b the situations s_1, s_2 are in the same stage. This represents that the probability of an individual having mild or severe symptoms does not depend on their treatment assignment - in other words, the treatment assignment and diagnosis are independent. This would be the case if the treatments have been suitably randomised. Additionally the situations s_3, s_5 are in the same stage, as well as s_4, s_6, which represents that the symptoms after one week are independent of the treatment type, conditional on the diagnosis severity.

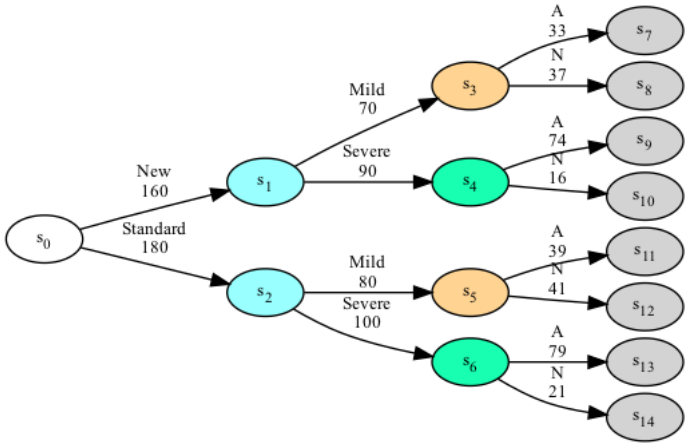
A *chain event graph* (CEG) is a more concise graphical representation of a ST model. All leaf vertices are combined into a single *sink* vertex and any two vertices with identical probability distributions for the remainder of the process are combined - such vertices are said to be in the same *position* (for more information on transforming a ST into a CEG see Shenvi and Smith, 2020). In Figure 1c the sink vertex is labelled w_∞ and the situations s_3, s_5 and s_4, s_6 have been merged into the two vertices w_2 and w_3. Additionally, the situations s_1, s_2 are in the same position so are combined into a single vertex w_1. Notice that the colours of the vertices in the CEG are inherited from the ST. In Figure 1c, the edges of the CEG have also been labelled with estimated probabilities.

ST models can be seen as a generalisation of Bayesian networks (BNs). For example, the ST in Figure 1b is equivalent to the BN with a single edge between the diagnosis severity and the symptoms. In fact, the class of ST models contains all possible BNs (Smith and Anderson, 2008). However, STs are a much richer class of models for two reasons:

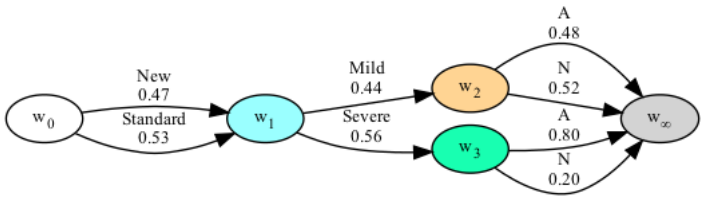
- STs are able to specify non-symmetric relationships between variables. For example, consider the adaptation of the ST in Figure 1b in which s_3, s_5 are still in the same stage but s_4, s_6 are no longer in the same stage. This represents that the two treatments perform equally for those with mild depression but differently for



(a) Event tree with observed counts



(b) Staged tree with observed counts



(c) Chain event graph with estimated probabilities

Fig. 1: Construction of a CEG for the single time point depression example.

those with severe depression. This relationship can be uncovered by a ST but not a BN. Although extensions of BNs representing non-symmetric types of independence have been proposed (e.g. Ankinakatte and Edwards, 2015; Jaeger et al., 2006; Pensar et al., 2015), these often lose the intuitiveness of BNs and no software to learn them from data is freely available.

- BNs require that the data be defined via a collection of variables - in STs this is referred to as stratified data. However, STs can also be used with more complicated forms of data - so called non-stratified data. Non-stratified data is associated to non-symmetric event trees and can occur due to structural zeros or processes where the set of possible future events depends on past events. In this paper we will focus on the stratified case but mention briefly the non-stratified case in Section 4.

2.2 Learning STs from data

There are two main approaches to using STs with data. The first assumes a fixed ST, based on expert or prior knowledge, and estimates transition probabilities under this ST. This helps to improve estimation because samples can be shared across all situations in the same stage which is particularly beneficial when sample sizes are small. For example, in Figure 1b, because s_1, s_2 are in the same stage, to estimate probabilities of mild or severe diagnosis we can use all samples from the data set, rather than calculating probabilities separately for those on the new and standard treatments.

The second approach is to perform model selection and learn the staging, and consequently non-symmetric types of independence, from data. Model selection of STs is often performed using Bayesian methods with a model prior over the space of STs and Dirichlet priors on the transition probabilities. The model space prior is often chosen to be uniform. A popular default choice for the Dirichlet priors is based on a score equivalence principle (see Collazo et al., 2018, for more details). However, in this paper we prefer setting all Dirichlet prior parameters equal to 1. Discussion about this is in Section 3.4. The model with highest posterior probability is then selected which can be done by comparing Bayes factors between models. However, due to the size of the space of ST models, it is often too computationally expensive to perform an exhaustive comparison of all models, hence search algorithms have been proposed to explore the space and approximate the highest posterior model. The first proposed algorithm, introduced in Freeman and Smith (2011b), is called the *agglomerative hierarchical clustering* (AHC) algorithm. The algorithm starts with a staged tree with each situation in its own stage and sequentially finds the two best stages to be merged, where best means highest Bayes factor. The algorithm continues until no increase in Bayes factor is found. Throughout this paper we will perform model selection with the AHC algorithm using the `cegpy` Python library (Walley et al., 2023).

More nuanced structural learning algorithms have since been developed (e.g Collazo and Smith, 2016; Leonelli and Varando, 2024; Silander and Leong, 2013). Most often, these search for the staged tree optimizing the model BIC (Görgen et al., 2022), using hill-climbing greedy algorithms. The R package `stagedtrees` (Carli et al., 2022) implements a wide array of these routines.

The two approaches of a fixed ST and model selection can be combined, as will be proposed in Section 3.2, by selecting the best ST using data-driven routines that respects a set of domain-based assumptions. This is analogous to blacklisting and whitelisting edges in BNs as for instance implemented in the `bnlearn` R package (Scutari, 2010).

A key theme of this paper is of a small sample size relative to the size of the event tree. In this case it is likely that some paths or situations in the event tree will have zero observations. In such cases, we propose a slight adaptation to the model selection process and the graphical representation of the selected CEG. First, all zero sample size situations are placed in the same stage (where edge labels allow for this). Note that these zero sample size stages are fundamentally different to the usual definition of a stage - they do not represent equality of probability distributions but instead the property of having no observed samples in the data set. Then, when performing model selection, we do not allow these zero sample size stages to be combined with any other stage. This can be achieved in the `cegpy` package using a combination of the `initial_staging` and `hyperstage` arguments. In the selected CEG, we represent the zero sample size stages by a square vertex coloured in grey and any edges with zero sample size with a grey dotted line.

Performing model selection and representing the CEG in this way has two advantages:

- Collecting all zero sample size situations together can greatly reduce the size of the resulting CEG and improve readability.
- Often the edges of a CEG are labelled with estimated probabilities. When estimated using Bayesian methods, this will usually be the maximum a posteriori estimate. However, in the case of zero samples, the posterior distribution is equal to the prior. By explicitly representing zero sample size situations in the graph, one can easily identify when estimated probabilities are just an artifact of the prior distributions rather than informed by data.

3 Longitudinal ST models

We now introduce our three proposed methods for modelling longitudinal data with STs. We will demonstrate the proposed methods on a longitudinal version of the depression example introduced in Section 2.

3.1 Full longitudinal ST

Longitudinal data fits naturally into a ST structure. This is because a temporal ordering of the events in the tree is often implicitly assumed - events that occur later in time appear later in the tree. Using a ST to model discrete longitudinal data is therefore just a matter of placing all events in a tree with the only modeling choice being the ordering of the variables that are not specified by the temporal ordering. Most commonly this will place time invariant covariates at the beginning of the event tree (Ankinakatte and Edwards, 2015) and longitudinal covariates before their associated outcome so

that the total ordering is $(Z, X_1, Y_1, \ldots, X_T, Y_T)$. Standard ST methodology can then be used for probability estimation or model selection.

With longitudinal data, situations associated with the same variable but at different times can be in the same stage. This can, for example, show when the probability distribution of a variable does not change over time. Such model selection is possible using the `cegpy` Python package in which any two situations with the same sample space can be selected to be in the same stage. However, the `stagedtrees` R package only allows situations associated to the same variable at the same time to be in the same stage so cannot learn such stagings.

Example 1 In the depression example from Section 2, symptoms are now assessed to be either normal (N) or abnormal (A) after 1, 2 and 4 weeks of treatment. The data set is summarised by the event tree in Figure 2 with edges additionally labelled with observed counts.

The highest posterior probability model found by the AHC algorithm is displayed in Figure 3. Situations related to outcome variables have been grouped into six stages from the singleton stage at w_{15} with highest probability of normal symptoms equal to 0.97 to the yellow stage with the lowest probability equal to 0.2. In particular, in week 1 there is no difference between the two treatments - those with mild depression have 0.51 probability of normal symptoms while those with severe depression only have 0.2 probability. The differences become evident in weeks 2 and 4. For those with severe depression, taking the standard treatment results in the probability of normal symptoms remaining low at 0.2 or 0.33 in week 2 and this only increases to 0.51 in week 4 for some individuals. However, taking the new drug results in the probability increasing to 0.51 in week 2 and increasing further to either 0.71 or 0.85 in week 4. Similar can be said for those with mild depression, with the new drug resulting in higher probabilities of normal symptoms in weeks 2 and 4.

This example shows how STs can be useful for analysing longitudinal data and the sort of conclusions that can be made. The main advantage of staged tree models is their ability to find context specific properties. This is even more powerful in longitudinal STs because situations related to the outcome at different times are able to be in the same stage. For example, we are able to see when the probability of normal symptoms remains unchanged over two different time points, despite the treatment.

However, some limitations of this approach are related to the size of the event tree and the sample size. As the number of paths in the graph increases (due to more time points, additional covariates or outcomes with larger sample space), interpretation of the resulting CEG can be challenging. Furthermore, the larger the tree, the larger the sample size required to ensure an adequate number of observations at each situation in the tree. This is already somewhat evident in the depression example where the situation s_{30} has only four observations.

3.2 Longitudinal ST with Markov assumptions

One way to utilise a small sample size more efficiently in staged tree models is to assume that certain situations are in the same stage i.e. they are assumed to have the same probabilities of the next event. Hence, the samples of these situations can be shared in the estimation of their probabilities. This can also be seen as reducing the

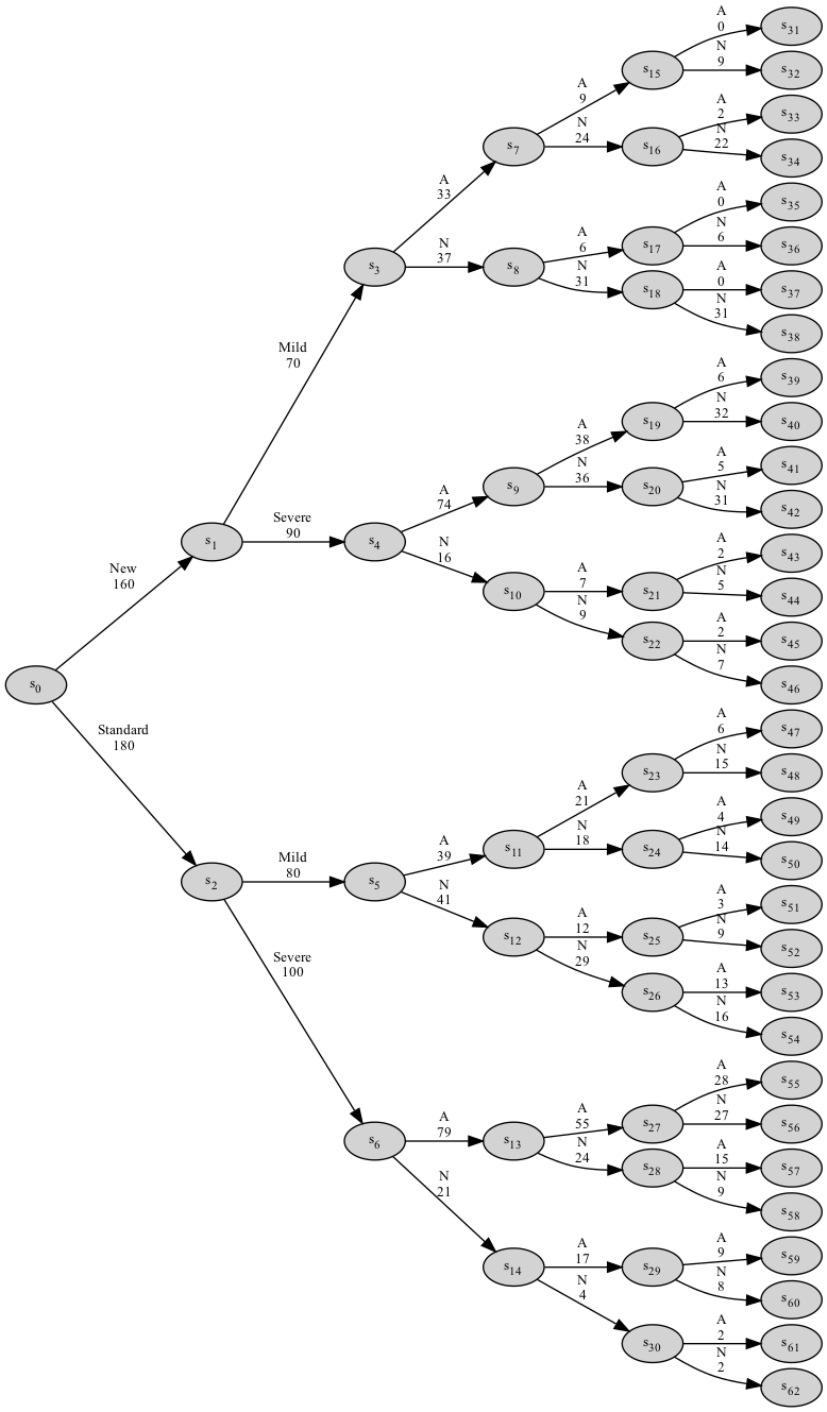


Fig. 2: Event tree for depression example. Edges are labelled with observed counts.

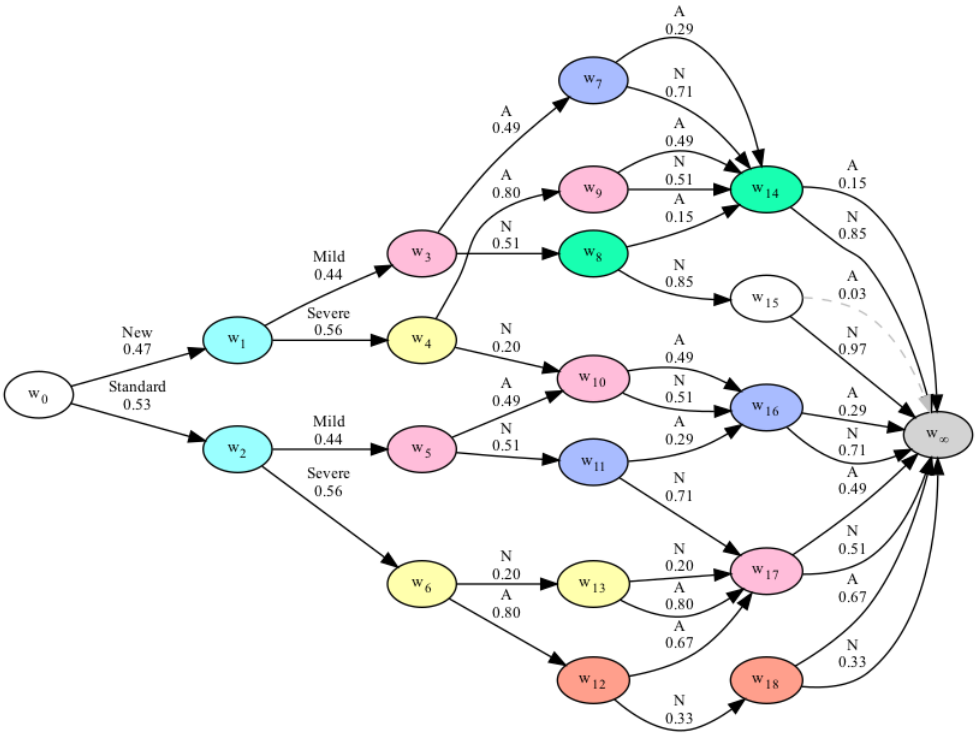
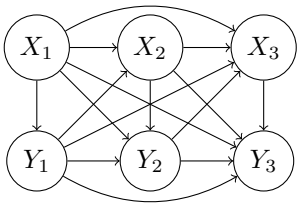


Fig. 3: CEG for depression example. Edges are labelled with estimated probabilities.

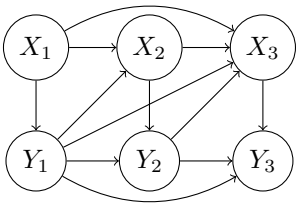
number of parameters in the model. After this initial staging is assumed, additional model selection can be performed. This strategy is adopted by Leonelli and Varando (2022) to limit the number of model parameters by enforcing an upper bound on the number of allowed dependencies. This procedure can be performed in the `cegpy` package using the `initial_staging` argument - a feature that has been added to the package for the purposes of this paper.

A simple way to specify which situations to place in the same stage is through conditional independence assumptions of the variables, which can be represented via the Markov assumptions of a directed acyclic graph (DAG). The choice of Markov assumptions could depend on the specific application and the aims of the analysis and be based on expert knowledge. Alternatively, one may follow the approach introduced in Barclay et al. (2013) by first selecting the DAG using data. A benefit of representing the assumptions by a DAG is that these can be easily converted to a ST using the algorithm of Varando et al. (2024) implemented in `stagedtrees`.

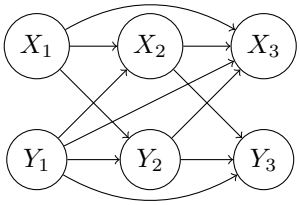
Although the specific Markov assumptions should be chosen based on the application, the longitudinal nature of the data supports some assumptions that might be relevant to many different applications. Some possibilities have been summarised in Figure 4.



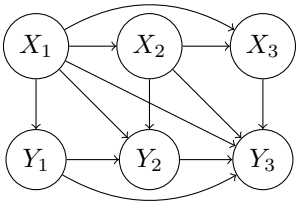
(a) Full model



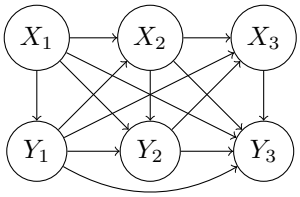
(b) Time dependence
$Y_t \perp\!\!\!\perp (X_1, \ldots, X_{t-1}) \mid (X_t, Y_1, \ldots, Y_{t-1})$



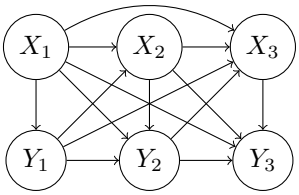
(c) Time lag dependence
$Y_t \perp\!\!\!\perp (X_1, \ldots, X_{t-2}) \mid (X_{t-1}, Y_1, \ldots, Y_{t-1})$



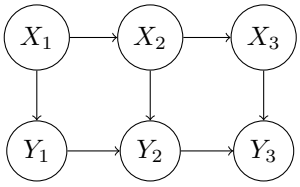
(d) Exogenous covariates
$X_t \perp\!\!\!\perp (Y_1, \ldots, Y_{t-1}) \mid (X_1, \ldots, X_{t-1})$



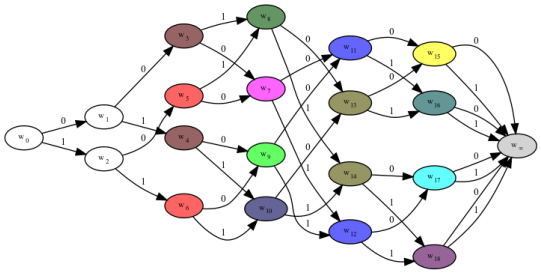
(e) Markov covariates
$X_t \perp\!\!\!\perp (X_1, \ldots, X_{t-2}) \mid (X_{t-1}, Y_1, \ldots, Y_t)$



(f) Markov outcome
$Y_t \perp\!\!\!\perp (Y_1, \ldots, Y_{t-2}) \mid (Y_{t-1}, X_1, \ldots, X_t)$



(g) Time dependence, exogenous covariates, Markov covariates and outcome



(h) CEG representation of the DAG in (g) with all binary variables

Fig. 4: Possible Markov assumptions and their DAGs for covariates and responses observed at three time points. Final graph shows the CEG with all binary variables associated to the DAG in (g).

A common assumption in regression models, for example the marginal model, is that the outcome at time t only depends on the time dependent covariates through those at time t (Figure 4b). Alternatively, if X_t and Y_t are measured simultaneously, the effect of the covariates might be lagged so that Y_t only depends on X_{t-1} (Figure 4c). Assumptions can also be made about how future covariates depend on past outcomes. Most simply, we might assume that they are conditionally independent or exogenous (Hernán et al., 2001) (Figure 4d). Further possible assumptions are that the outcomes or the covariates follow Markov processes (Figures 4e and 4f).

Combining a number of these assumptions can greatly simplify the staged tree. For example, Figure 4g shows a DAG for three time points representing the assumptions of time dependence, exogenous covariates and Markov covariates and outcomes, alongside the equivalent CEG representation of these assumptions with binary variables in Figure 4h. This CEG, for example, has combined the 32 situations associated to the final outcome Y_3 into only 4 stages.

Example 2 A simple assumption we could make in the depression example is that the outcomes follow a Markov process. This corresponds to placing certain vertices in the final level of the tree in the same stage. Therefore the sample size can be shared across such vertices. For example, returning to Figure 2, the situation s_{30} has a sample size of only 4. But with the Markov assumption, s_{30} is in the same stage as s_{28} and their combined sample size is 28.

With this Markov assumption, the selected CEG (shown in Figure 5) has been simplified somewhat with each of the four covariate combinations being separated. Furthermore, for patients using the new drug or using the standard treatment with mild symptoms, the outcome in week 4 is independent of both previous outcomes.

3.3 Marginal longitudinal STs

In the previous section, Markov assumptions were used to improve inferences in cases where sample sizes are too small for the given event tree. However, when the event tree is large, the graphical representation of the ST or CEG can be hard to interpret making inferences challenging to read. Instead it may be useful to consider certain marginal distributions which generate event trees of smaller size making the resulting CEG easier to interpret. It also reduces the number of parameters in the model, improving estimation is cases of small sample sizes.

The choice of marginal distributions depends on the application and the aims of the analysis. For example, if one is interested in the marginal distributions at each time point we might fit different STs for each (Z, X_t, Y_t), $t = 1, \ldots, T$. Alternatively, if one is interested in the long term effects of the covariates on the final outcome, we might instead consider the margins (Z, X_t, Y_t, Y_T), $t = 1, \ldots, T-1$.

A limitation to this approach is that it does not directly use the longitudinal nature of the data and therefore misses out on its potential benefits. For example, the marginal regression model is able to utilise the whole data when estimating the marginal probabilities by specifying a covariance structure between the repeated observations. While doing this for STs is more difficult - unlike in regression models, a ST explicitly models the covariates and so a covariance structure is required for all timed covariates and

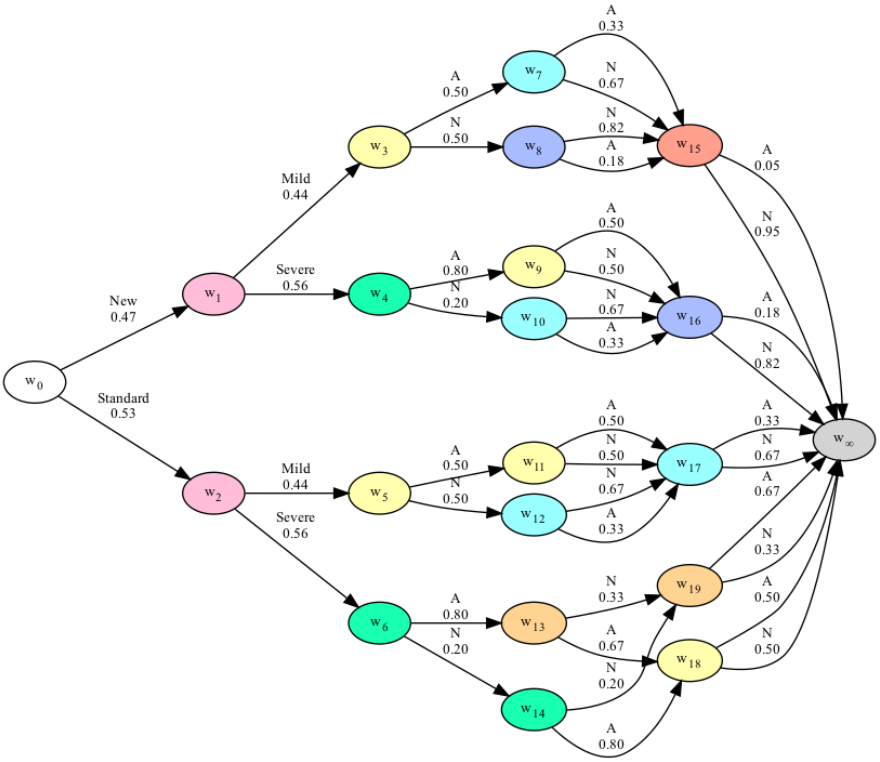


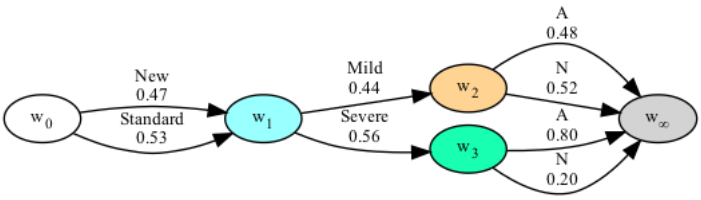
Fig. 5: CEG for depression example with Markov process outcomes assumption.

outcomes - there are some alternative ways to incorporate the longitudinal nature of the data into STs.

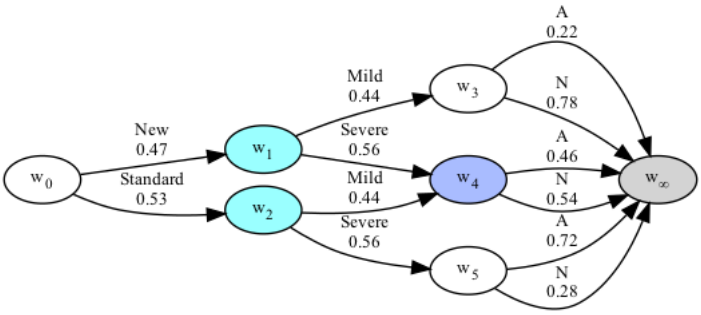
Often in longitudinal data the covariates and outcomes at different times share the same sample space and so the event trees for each of the margins share the same topology. One can therefore conduct further analysis by comparing the stagings of the selected STs for the different margins. This can give an idea of how the relationship between the outcome and covariates changes over time.

When the outcome variable is numeric, another way to incorporate the longitudinal nature of the data is to redefine the outcome variables as the difference between the outcome at different time points. This was the approach taken by, for example, Scutari et al. (2017) when using BNs to analyse longitudinal data. This focuses the analysis on how the outcome changes over time, rather than on its numeric value. An example of this will be given in Section 5, but first we return to the depression example.

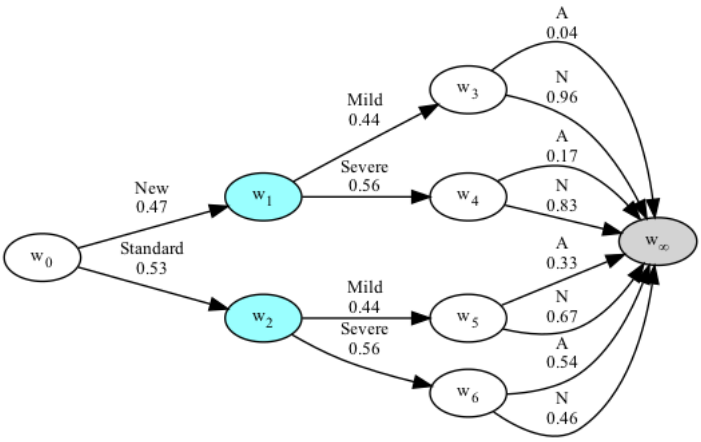
Example 3 We consider the marginal distribution of the covariates with each of the three timed outcomes individually to uncover the marginal effect of the two treatments at different times. Note that now the smallest sample size for a situation is 70. The selected CEGs for the three margins are in Figure 6.



(a) Week 1



(b) Week 2



(c) Week 4

Fig. 6: Marginal CEGs for depression example.

In week 1, the new drug and standard treatment are placed in the same position - this means that in the model there is no difference between the treatments at this time. In week 2 the treatments are no longer in the same position. However, the new drug with severe symptoms is in the same stage at the standard treatment with mild symptoms, both with a 0.54 probability of normal symptoms. The difference between the two treatments is even greater in week 4 when all situations before the outcome are in different stages. The new

treatment leads to higher probability of normal symptoms in week 4 no matter the diagnosis severity.

3.4 Prior parameters

Bayesian model selection requires specification of a prior distribution for the transition probabilities. The conjugate prior places independent Dirichlet priors on the transition probabilities for each situation. This is the most popular choice of prior because it allows for simple computation and is the only form of prior currently implemented in the `cegpy` package. However, there is still flexibility in the choice of parameters for the Dirichlet priors.

The default choice of parameters in the `cegpy` package is based on a score equivalence principle and is calculated by propagating an equivalent sample size along each edge. The rationale behind this prior is that when changing the order of the event tree, it is possible to have different STs that are equivalent in terms of the assumptions they place on the probability distribution. The score equivalence prior ensures that any two equivalent STs have the same posterior probability (Cowell and Smith, 2014; Hughes et al., 2022).

With longitudinal data the principal of score equivalence is less relevant because the underlying event tree can be fixed according to the temporal ordering of the data. When the event tree is fixed, all ST models are distinct and so score equivalence is not necessary. Instead, an additional consideration particularly relevant for longitudinal STs is that situations at different levels of the tree can be places in the same stage. However, the score equivalence prior places different priors on situations at different levels of the tree. This can create somewhat of an imbalance performing model selection.

For this reason, we instead choose every Dirichlet prior parameter to be equal to a common value. For the examples in this paper we choose this common value to be 1 so that all binary probabilities are uniform under the prior. However, larger or smaller values can be chosen depending on the strength of prior that is wanted - smaller values tend to result in simpler models where more situations are joined in the same stage, while larger values join less situations.

Further consideration must also be made for longitudinal STs with Markov assumptions. The Markov assumptions join certain situations into the same stage a priori which has the effect of summing the Dirichlet parameters in the prior for the stage transition probabilities. When the Markov assumptions combine many situations into the same stage, this can result in the prior being unexpectedly strong. For this reason, we propose dividing the prior parameters by the number of situations joined together by the Markov assumptions. This ensures that all prior parameters are the same after the Markov assumptions have been applied.

3.5 Robustness

A key reason for proposing the Markov assumptions for longitudinal STs is to make better use of data with small sample sizes. It is therefore important to consider how robust these methods are to small sample sizes in terms of both probability estimation

and model selection. In cases where the chosen Markov assumptions are in reality true, this seems obvious - the inclusion of these correct assumptions allows more samples to be shared across different situations, increasing the sample size and improving estimation of the transition probabilities. But what about when the chosen Markov assumptions are not true?

In the depression example there is some evidence that the chosen Markov assumption is not strictly true. For example, looking at Figure 2, the Markov assumptions joins situations s_{24} and s_{26} into the same stage. However, there are enough samples for these two situations to conclude that they should not be in the same stage - the MLEs for the probability of normal symptoms are 0.78 and 0.55 respectively. We continue by investigating how the Markov assumption affects probability estimation when using smaller sample sizes.

We begin with the full longitudinal ST displayed in Figure 3 fit on the full data set. We then remove some proportion of the data and fit both a full longitudinal ST and longitudinal ST with Markov assumptions to this new smaller data set. We then measure the difference in probability estimates in comparison to the model fit on the full data. As a metric we use the total variation distance of transition probabilities averaged over the situations. A smaller distance reflects a smaller change in the estimated probabilities and therefore more robustness to smaller sample sizes.

This was repeated 1000 times for a range of proportions of data removed. The results are displayed in Figure 7. We see that when only a small proportion of the data is removed, the Markov assumptions mean that the probability estimates are further from those estimated on the full data. This is to be expected because, as discussed, the Markov assumptions are not fully supported by the data and there is still sufficient sample size to adequately estimate the transition probabilities. However, as more data is removed the Markov assumptions help the estimated probabilities remain closer to those from the full data. This shows that the addition of Markov assumptions, even if the Markov assumptions are not strictly true, can help the probability estimates to be less sensitive to removing data points and more robust with small sample sizes. More details on the experiment design and additional results can be found in Appendix A.

4 Comparison to regression methods and benefits of STs

In this section we compare the proposed ST methods to regression based methods and discuss some of the benefits of using STs for longitudinal data. We start by continuing the depression example, comparing the estimated probabilities obtained by the ST methods in the previous sections to a marginal regression model. We then give a more general discussion on the differences between ST models and regression models, focusing on the different assumptions they make and the conclusions at which they are able to arrive. To begin we give a brief introduction to three standard regression based approaches for modelling longitudinal data with linear models - the marginal model, transition model and random effects model (see Diggle et al., 2002, Chapters 7 - 10). The primary difference between these three models is in how they characterise the dependence between the outcomes at different time points.

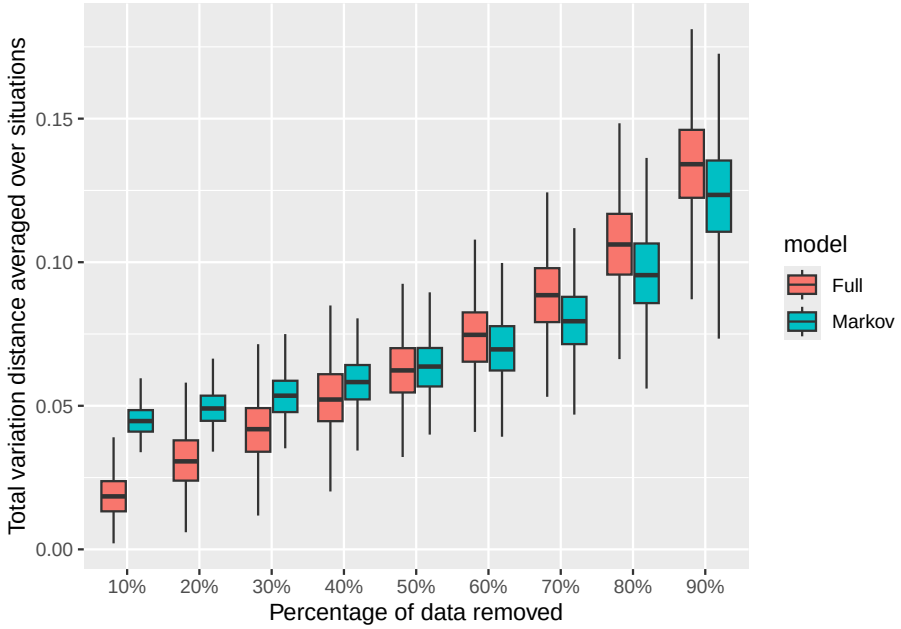


Fig. 7: Boxplots of the average situation total variation distance to the full longitudinal ST for models fit on subsets of the data.

The marginal model addresses the regression and longitudinal dependence separately. A common regression model is assumed over each time point with outcome Y_t and covariates Z, X_t. Although the model is assumed to be the same at each time point, differences between the time points can be expressed through the inclusion of time t as a covariate, as well as interaction terms between t and the other covariates. Defining the likelihood function also requires a model for the dependence between the outcomes $Y_1, \ldots, Y_T$. With discrete outcomes, this is a complex task which involves many nuisance parameters. Instead, a generalised estimating equations approach only requires a model for the pairwise correlations. Common choices for this correlation model are an exchangeable structure where $\text{Corr}(Y_s, Y_t)$ is identical (but unknown) for all s, t or an autoregressive structure.

The transition model captures both the effect of the covariates on the outcome and the time dependence within the same logistic regression. This is achieved by including functions of the previous outcomes $Y_1, \ldots, Y_{t-1}$ as predictors in the logistic regression for Y_t. This allows for a more nuanced description of the time dependence than in the marginal model, however additional care must be taken in the interpretation of the regression coefficients because their value depends the chosen functions of previous outcomes.

The random effects model takes a different approach by allowing some regression coefficients β_i to vary across individuals i. It is then assumed that any dependence

between the longitudinal outcomes is captured by β_i - that is, $Y_1, \ldots, Y_T$ are conditionally independent given β_i. Additionally, it is required to specify a distribution for the coefficients β_i - a common choice is a zero mean Gaussian distribution. This approach is useful because it is allows inference on the individual level rather than just the population level.

4.1 Depression example

The depression data set was analysed using a marginal regression model in Agresti (2012), Chapter 9. The three time points are labelled $t = 0, 1, 2$. The outcome variables are Y_0, Y_1, Y_2 with $Y_t = 1$ denoting normal symptoms and $Y_t = 0$ abnormal symptoms. The treatment allocation is denoted by $d = 0$ for the standard treatment and $d = 1$ for the new drug, and the diagnosis severity by $s = 0$ for mild and $s = 1$ for severe. The logistic regression model used by Agresti (2012) was

$$\text{logit}\left(P(Y_t = 1 \mid s, d)\right) = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt$$

Additionally, an exchangeable correlation structure is used so that $\text{Corr}(Y_s, Y_t) = \rho$ for all pairs s, t. Parameter estimates are obtained using the generalised estimating equation method.

The fitted correlation between different times was very weak with estimated $\hat{\rho} = -0.003$. This is somewhat reflected in the CEG in Figure 3, however the CEG also highlights that this independence is context specific - it only occurs for certain combinations of the covariates and previous outcomes. For example, Y_1 is independent of Y_0 when $d = 1$ and $s = 1$. However, they are not independent for any of the other combinations of covariates.

The general conclusions from the estimated regression coefficients also match the conclusions from the ST models. The estimated time effects are $\hat{\beta}_3 = 0.482$ for the standard treatment and $\hat{\beta}_3 + \hat{\beta}_4 = 1.500$ for the new drug. The Wald test for $\beta_4 = 0$ has p-value < 0.0001 demonstrating strong evidence for a faster improvement in symptoms when using the new drug. Additionally, the estimated effect of the new drug at time $t = 0$ is $\hat{\beta}_2 = -0.060$ - a very small effect demonstrating an insignificant difference between the treatments in the first week. Both of these conclusions match those of all three ST methods.

Estimated marginal probabilities for the outcome at each time point conditional on the covariates can be found in Table 1 for each of the three ST methods and regression model. While these probabilities are generally very similar, there are some notable differences. Between the ST methods, differences in estimated probabilities occur due to different stagings in the selected models. For example, for $d = 0, s = 0$ the estimate of the probability of $Y_2 = 1$ is 0.61 under the full ST, but only 0.54 under the marginal ST. This is because the marginal ST includes this situation in the same stage as $d = 1, s = 1$.

On the other hand, the regression model can obtain different probability estimates due to the linear assumptions in the model. In particular, it assumes that the probabilities change linearly over time with respect to the logit link function. This results in the estimated probabilities for $d = 0, s = 0$ of 0.49, 0.61 and 0.72 for weeks 1, 2 and 4

Covariates	Model	Week 1	Week 2	Week 4
New drug, severe	Full ST	0.20	0.51	0.82
	Markov ST	0.20	0.53	0.82
	Marginal STs	0.20	0.54	0.83
	Regression	0.20	0.52	0.83
New drug, mild	Full ST	0.51	0.78	0.90
	Markov ST	0.50	0.75	0.95
	Marginal STs	0.52	0.78	0.96
	Regression	0.48	0.80	0.95
Standard treatment, severe	Full ST	0.20	0.30	0.46
	Markov ST	0.20	0.30	0.45
	Marginal STs	0.20	0.28	0.46
	Regression	0.21	0.30	0.41
Standard treatment, mild	Full ST	0.51	0.61	0.64
	Markov ST	0.50	0.59	0.67
	Marginal STs	0.52	0.54	0.67
	Regression	0.49	0.61	0.72

Table 1: Estimated marginal probabilities of normal symptoms for each time point conditional on covariate values

respectively. In comparison, the full ST model estimates these probabilities as 0.51, 0.61 and 0.64. Notice the large difference in estimated probabilities at week 4. Free from the linearity assumption, the ST method is able to estimate a lower probability demonstrating diminishing returns from the standard treatment over time.

The marginal STs and regression model are only able to estimate marginal probabilities of the outcome at each time. For the regression model, this is because the marginal probabilities and pairwise correlations do not specify the joint distribution of the outcomes. On the other hand, the full ST and Markov ST methods are able to estimate the full joint probabilities. This enables slightly more nuanced observations. For example, for $d = 1, s = 1$, all methods estimate the probability of normal symptoms at week 4 to be between 0.82 and 0.83. However, with the full ST model this probability is estimated to be 0.85 if the patient had abnormal symptoms at week 1 but only 0.71 if they had normal symptoms at week 1. Such context specific information could be used to reassure a patient that was worried after continuing to experience abnormal symptoms after one week of treatment.

4.2 Model assumptions

The full ST described in Section 3.1 makes no assumptions about the data beyond the sample space described by the event tree. This lack of assumptions can be a benefit because it makes it applicable to any appropriate data set and the conclusions are not dependent on any assumptions. However, as has been mentioned previously, this lack of assumptions requires a large sample size relative to the size of the event tree in order for probability estimation and model selection to be performed reliably. This is because the full joint probability distribution needs to be estimated and, as the event tree gets larger, this joint distribution involves many parameters. In situations where there is not sufficient data, one can reach conclusions more supported by data

by making additional assumptions - although the conclusions are now dependent on these assumptions being true.

Regression based models can make a number of assumptions depending on the approach taken. Common to all regression methods is that the outcome probabilities are linear functions of the covariates (with respect to a chosen link function). While this assumption can be relaxed slightly by including interaction terms between the covariates, it still results in the effect of the regression coefficients being additive.

Additionally, regression based methods make assumptions about the dependence between the outcome at different time points. The marginal model does this separately from the regression by assuming some correlation structure between the outcomes at different times. The transition model includes the previous outcomes in the regression resulting in an additive effect. The random effects model assumes independence of the outcomes conditional on some individual level parameter. Depending on the application, these assumptions may be more or less appropriate.

In Section 3.2 we suggested the use of ST models with Markov assumptions given by a DAG which specify conditional independences between the variables. However, beyond these conditional independences, no additional restrictions are placed on the model. The choice of Markov assumptions is left to the practitioner, allowing the assumptions to be chosen to best fit the specific application. Additionally, due to the DAG respresentation, these assumptions are easy to interpret and can be verified either using expert judgement or testing the conditional independences using data.

Like the full ST model, the marginal ST models of Section 3.3 make no assumptions about the data. However, they only model the marginal distributions rather than the full joint distribution so the conclusions one can draw are weaker. However, when combined with certain (conditional) independence assumptions, one could construct the full joint probability distribution. For example, the marginal STs for each individual time point (Z, X_t, Y_t), $t = 1, \ldots, T$ combined with the assumption that the time points are independent given Z results in the full joint probability distribution.

4.3 Model interpretation

The primary interest of regression based models is the effect of the covariates on the outcome. As such, they only model the conditional distribution of the outcome given the covariates. One is able to estimate these conditional distributions via the estimates for the regression coefficients. One can also use these regression coefficients to infer the effect of the covariates on the outcome. In the marginal model, interpretation of the regression coefficients is simple and analogous to a standard logistic regression. However, in the transitions model and random effects model this interpretation is more complicated since the value of the regression coefficients depends on other modelling choices (Zeger and Liang, 1992).

By including interaction terms in the regression between the covariates and time, one is able to determine how the effect of a covariate changes over time. However, this change is restricted to be linear (and therefore monotonic) over time. By testing for coefficients being equal to zero (either using p-values or more sophisticated model selection methods), one can also find which covariates have no effect on the outcome, or which effects do not change over time.

One can also use regression methods to infer the dependence between the different timed outcomes. How this is done depends on the specific method - the marginal model estimates a common covariance function between all outcomes while the transition model interprets the dependence through additional regression coefficients.

The conclusions made by the proposed ST methods are twofold. First there is the model selection step in which the ST with highest posterior probability that best describes the data is searched for. While there are often many models with high posterior probability (Strong and Smith, 2022), the selected model can still provide a useful interpretation of the data. For example, by identifying situations that are in the same stage, one can find different combinations of the covariates which have the same effect on the outcome in terms of the probability distribution.

Second, given the selected ST, transition probabilities are estimated. In this regard ST methods go one step further than regression methods because they explicitly model the covariates. This means that conclusions are not limited to the outcome variable but can also be made about the covariates - the dependence between the covariates and how they change over time. This can be particularly useful in a longitudinal study because if there is a covariate which has a strong effect on the outcome, it would also be important to understand how the value of this covariate changes over time.

4.4 Other benefits of STs

Aside from the differences discussed above, ST models also enjoy some additional benefits in the modelling of longitudinal data.

A key benefit of ST models, not limited to longitudinal data, is the graphical representation. The CEG provides a visual aid for interpreting the conclusions of the model selection which can be understood without expertise in how the model works. The event tree also provides a useful visualisation tool when the edges of the event tree are labelled with the number of observations in the data set. Using this, one can easily identify situations in the tree that have few or zero observations. This is important because any conclusions or probability estimation made about such small sample size situations will be based primarily on the modelling assumptions. In a ST modelling process, such limitations of the data set would become obvious when drawing the event tree and can be made evident in the selected CEG by using the new representation suggested at the end of Section 2. However, in standard regression modelling this would be harder to identify without careful exploratory analysis. In particular, the linear assumptions of regression models mean that strong inferences can be made about situations with few observations.

STs can be updated with only a partial observation of the variables (Freeman and Smith, 2011b). This is useful because data points which only observe the variables up to a certain time point - either due to dropout or because the individual has not completed all time points - can still be used to update probabilities.

In this paper we have distinguished the outcome variable from the covariates. Such a distinction is useful when the focus of the analysis is on the effect of the covariates on the outcome, as is the case in regression models. However, this distinction is not necessary for ST models allowing for a more general modelling approach. This can be

useful when one is instead interested in the distribution of all observed variables. Each of the three ST methods in this paper remain valid in this case.

In fact STs are able to generalise further. So far in this paper we have focused on stratified event trees in which each layer of the tree can be associated to a variable. However, a key benefit of ST models is that they do not require a variable based representation and can model asymmetric or *non-stratified* processes. This is useful in data sets which contain structural zeros - when some sequence of events occurs with probability zero - or when the sample space of an event depends on the previous events. This might occur in a longitudinal data set when what is measured at time t depends on the observed values at the previous times.

5 Dentistry data example

We now demonstrate the proposed ST methods on a longitudinal data set examining the prevalence of caries (also called tooth decay or cavities) in children. Initially a cross-sectional study was conducted and analysed by Ugolini et al. (2018) and, based on the findings of this study, a further longitudinal study was carried out by the same authors. An initial version of this longitudinal study with three time points was analysed by Ugolini et al. (2023) using an undirected graphical model. Here we have access to an extended version of this data set with four time points with time-dependent variables measured at ages 3, 4, 5 and 7. We denote these four time points by $t = 1, 2, 3, 4$.

In Ugolini et al. (2023), the selected undirected graphical model found a direct relationship between a child's oral hygiene and the prevalence of caries. There was also an edge between the length of breastfeeding time and oral hygiene, however the prevalence of caries was found to be independent of breastfeeding time conditional on oral hygiene. In this graphical model, repeated observations were summarised into a single variable and so it was more similar to a cross-sectional or marginal analysis. We aim to investigate the relationship between these three variables over the whole time series.

The data set contains the following variables:

- Breastfeeding time Z. A time invariant covariate indicating short breastfeeding time of 0-9 months (including no breastfeeding) or long breastfeeding time of 10+ months. The cutoff of 9 months was chosen to obtain two balanced classes.
- Oral hygiene $X_1, \ldots, X_4$. Longitudinal covariates giving a clinical assessment of whether the child's oral hygiene was 'adequate' or 'inadequate' measured at all four time points $t = 1, \ldots, 4$.
- Change in caries incidences Y_{ij}. At each age the total number of caries incidences (the sum of cavities and fillings) was recorded, denoted by $I_1, \ldots, I_4$. As the longitudinal outcome variables, we are interested in whether the number of incidences increases or not between different ages. These are denoted by the variables Y_{ij}, $i, j = 1, \ldots, 4$, $i < j$ where Y_{ij} is 'increasing' if $I_j - I_i > 0$ and is 'decreasing' if $I_j - I_i \leq 0$ (no change is included in the decreasing category). At the first recorded time point at age 3, the outcome is instead denoted by either Y_1 or Y_{01} and is 'caries' if $I_1 > 0$ and 'no caries' if $I_1 = 0$.

The data set contains 277 participants of which 237 have full observation of all variables (missing values were due to a seasonal flu and so can be considered missing at random). Modelling this data with a full ST model might consider the nine binary variables $(Z, X_1, Y_1, X_2, Y_{12}, X_3, Y_{23}, X_4, Y_{34})$. The corresponding event tree has 512 paths and with only 237 full observations, the sample size is not sufficient to reliably estimate the full joint probability. In fact, there are only observations for 37 of the 512 paths and 135 of the observations are on the two paths with all adequate hygiene and no caries. This indicates that modelling the data with the full longitudinal ST with no additional assumptions is not suitable.

Instead, as suggested in Section 3.3, we first consider certain marginal distributions of the data.

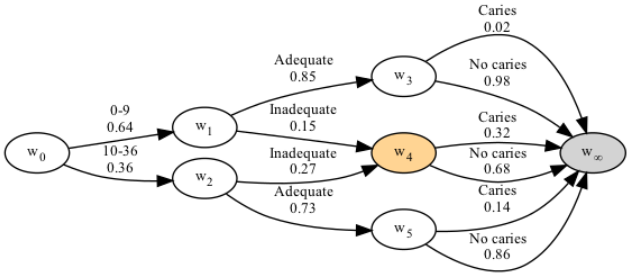
5.1 Marginal STs

The first marginal distributions we investigate are for each individual time point $(Z, X_t, Y_{t-1,t})$. The selected CEGs are in Figure 8. At ages 3 and 4, a long breastfeeding time is associated with a higher probability of inadequate hygiene and, in some cases, a higher probability of increasing caries incidences. However, at ages 5 and 7 breastfeeding time is marginally independent of oral hygiene and incidence change.

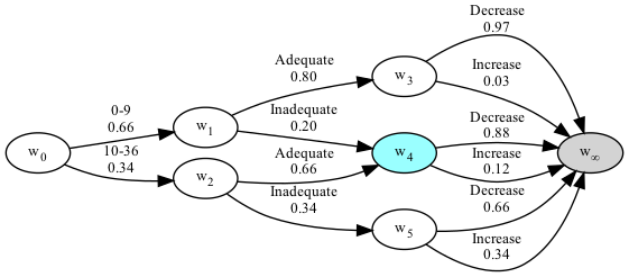
For a longer term effect of the covariates, the change in incidences to the final time point can be included at the end of each marginal CEG $(Z, X_t, Y_{t-1,t}, Y_{t7})$. The selected CEGs are in Figure 9. In all selected CEGs, paths with inadequate hygiene lead to situations with higher probability of increasing incidences at age 7, while paths with adequate hygiene generally lead to lower probability. However, for ages 3 and 5 the path with short breastfeeding time, adequate hygiene and increasing incidences arrives at a situation with a high probability of increasing incidences at age 7.

In the first marginal CEGs of Figure 8, we found evidence that inadequate hygiene is a risk factor for caries. Additionally, there is evidence that at younger ages a long breastfeeding time can increase the probability of inadequate hygiene, indirectly leading to higher risk of caries. There is also slightly weaker evidence that long breastfeeding time is a direct risk factor for caries at younger ages. However, in the marginal CEGs of Figure 9, it was found that individuals that have increasing caries incidences despite having neither of these two risk factors (i.e. short breastfeeding time and adequate hygiene) have a high probability of increasing incidences in the future. This perhaps suggests the presence of a further risk factor not included in this data set - although it should be noted that this is based on relatively few observations.

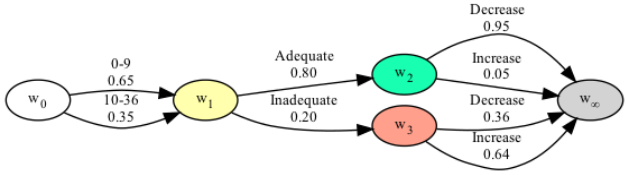
The marginal CEGs of Figure 9 further show that inadequate hygiene at younger ages is associated with a higher probability of increasing incidences at older ages. Within this data set, there are two likely reasons for this. One is that hygiene at different ages are highly correlated, i.e. inadequate hygiene at a younger age is associated to inadequate hygiene at older ages which is in turn a risk factor for caries at older ages. The other is that inadequate hygiene at younger ages is a direct risk factor for caries at older ages. To investigate this further we consider the marginal CEG for the increase in incidences at age 7 with all covariates $(Z, X_1, X_2, X_3, X_4, Y_{34})$. The selected CEG is in Figure 10. In this CEG, there is one stage with low probability of



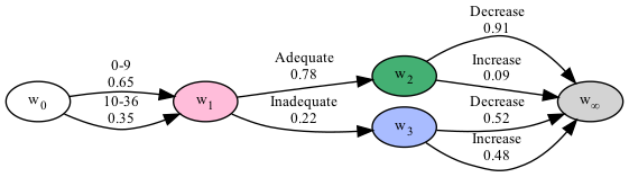
(a) Age 3



(b) Age 4



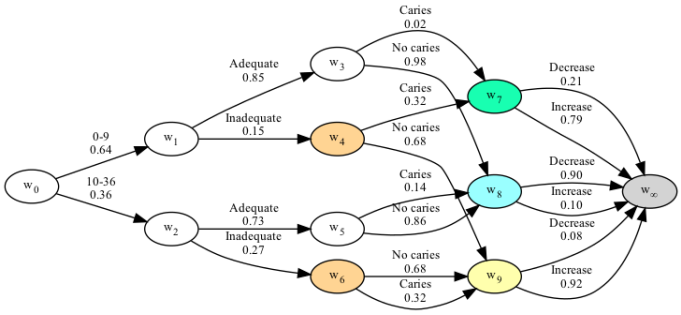
(c) Age 5



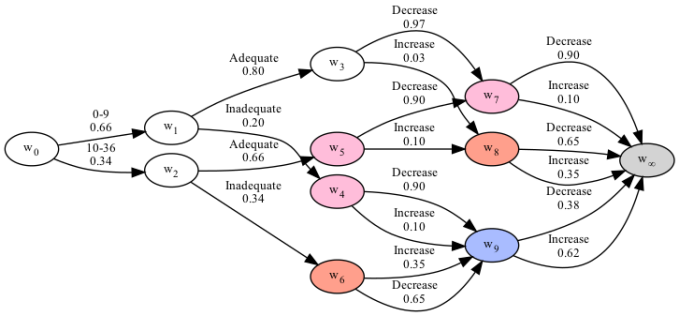
(d) Age 7

Fig. 8: Marginal CEGs at each time point $(Z, X_t, Y_{t-1,t})$.

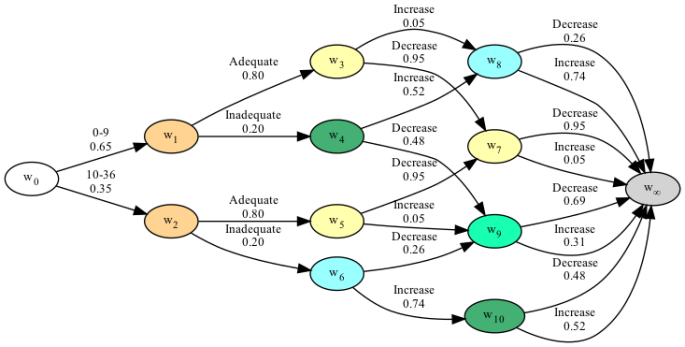
increasing incidences - w_{26}. All paths leading to this stage has either 0 or 1 occurrence of inadequate hygiene throughout the time series. On the other hand, all paths leading to the stage with highest probability of increasing incidences, w_{29}, have 3 or 4 occurrences of inadequate hygiene. A similar phenomenon was also observed for age 5 but this CEG is omitted for brevity. Based on this, we make the hypothesis that the number of occurrences of inadequate hygiene is a risk factor for caries, rather than the



(a) Age 3



(b) Age 4



(c) Age 5

Fig. 9: Marginal CEGs for each time point with change to age 7 $(Z, X_t, Y_{t-1,t}, Y_{t7})$.

exact time that they occur. Note that this is a context specific property which could not be discovered directly from a BN.

5.2 ST with Markov assumptions

We now select a ST for the full data set $(Z, X_1, Y_1, X_2, Y_{12}, X_3, Y_{23}, X_4, Y_{34})$ with additional Markov assumptions (Section 3.3). These assumptions are in part informed by

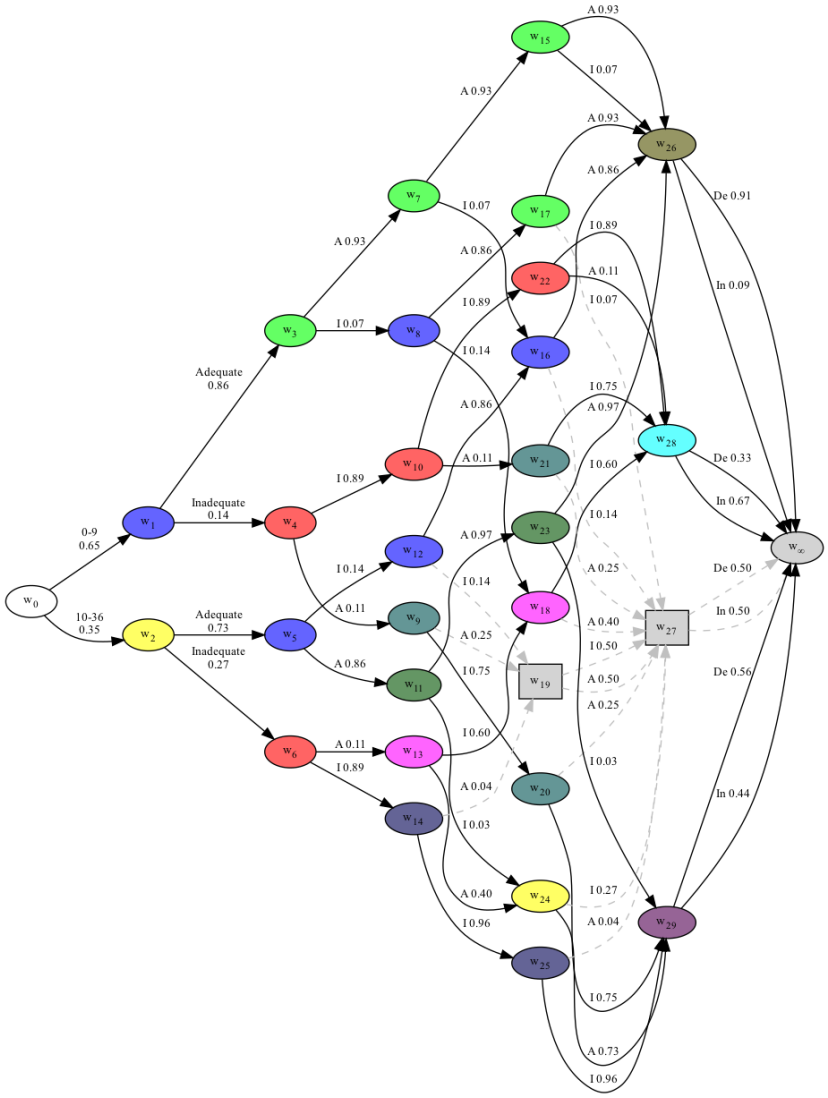
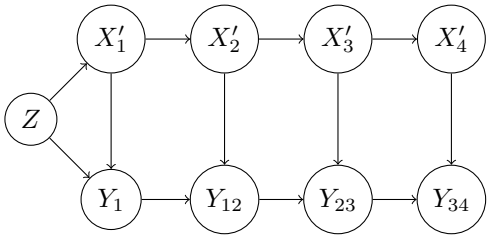
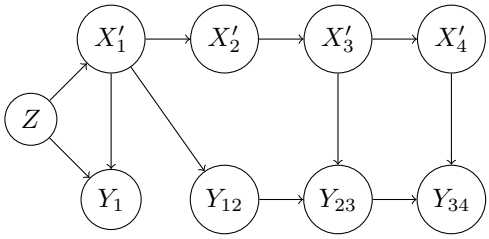


Fig. 10: Marginal CEG for incidence change at age 7 with all past covariates $(Z, X_1, X_2, X_3, X_4, Y_{34})$. Some edge labels are abbreviated: Adequate (A), Inadequate (I), Increase (In), Decrease (De).

the conclusions of the marginal STs in Section 5.1. Thus, to specify the assumptions we consider a DAG using redefined hygiene variables that count the number of inadequate occurrences - $X'_t = \sum_{i=1}^{t} X_i$. We refer to these as the hygiene sums. Note that any BN using the hygiene sums can still be expressed using a ST model on the original hygiene variables.



(a) DAG representing chosen assumptions in ST model



(b) DAG selected by data

Fig. 11: DAGs for dentistry data with redefined hygiene variables $X'_t = \sum_{i=1}^{t} X_i$.

Our chosen Markov assumptions are represented in the DAG in Figure 11a and can be summarised as:

- Ages 4, 5 and 7 are conditionally independent of breastfeeding time given age 3.
- Hygiene sums follow a Markov process and are exogenous.
- Incidences follow a Markov process and only directly depend on hygiene through the sum at the corresponding time.

To further validate these assumptions, we select a BN model from the data using the R package `bnlearn` (Scutari, 2010). We first specified that the direction of edges in the BN should respect the chronology of the data. Then, to select between different models we used the Akaike Information Criterion and used Tabu search to explore the model space for the highest scoring model. The selected BN is in Figure 11b. The selected BN closely matches our chosen BN, lending strength to the assumptions that were made.

The chosen assumptions can now be specified in a ST model and additional model selection conducted. The selected CEG is in Figure 12. Despite there being no observations on the majority of paths through the tree, with the benefit of the Markov assumptions there are only 4 situations that have zero sample size which are all associated to the outcome at age 7 (these have been combined into a single zero sample size vertex labelled w_{40}). The CEG can be used to calculate estimated joint probabilities or to track a child through the CEG and provide estimated probabilities for future caries incidences.

The addition of the Markov assumptions has allowed probability estimation for situations that have few or zero observations. For example, in the selected CEG w_{33} has no observations. However, because it is in the same stage as w_{32}, we are able to estimate the probability of adequate hygiene as 0.96.

6 Discussion

In this paper we have discussed various approaches for using STs to model discrete longitudinal data. Each of these take advantage of existing methodology for ST model selection and probability estimation and so can be easily implemented - namely by using the Python package `cegpy` or the R package `stagedtrees`.

A main advantage of the proposed ST approaches are the simple assumptions that they make. Both the full longitudinal ST and marginal longitudinal STs do not make any assumptions about the data. The longitudinal ST with Markov assumptions only makes assumptions about the conditional independence of the variables (although more complicated assumptions such as context specific independencies are also possible). Such assumptions are easy to interpret and have a visual representation through a DAG. Additionally, the assumptions can be verified through standard methods for BN model selection. This is in contrast to regression methods whose assumptions are often less explicit and harder to interpret.

Another contribution of this paper is the new representation of a CEG for zero sample size situations. Creating a separate stage for zero sample size situations allows for a more compact and readable representation of the CEG while also providing additional information about the data set and estimated probabilities. Specifically, it highlights any situations with zero sample size so that one knows that probability estimation for that situation is based purely on the prior distribution and not on data. This could also be used to inform new data collection.

Future work in this area could focus on the development of methodology, for example structural learning algorithms, that are tailored to longitudinal data taking into account the dependence between different time points. One possibility for this is to define an event tree for the whole data set but with a different branch of the tree for each time point. Another is to assume that the marginal ST remains the same at each time point so that one can use the whole data to select this marginal ST. An issue with these approaches, similar to the marginal regression model, is that observations are not independent due to the dependence between different time points for the same individual. Defining the likelihood function therefore requires specification of this dependence between different time points. This is even more complicated for ST models because the covariates are explicitly modelled - hence the dependence must be specified for all timed covariates as well as the outcome. However, marginal regression models simplify this process significantly through the use of generalised estimating equations which don't require the full likelihood. An interesting line of future research would be in the use of the generalised estimating equations for ST models.

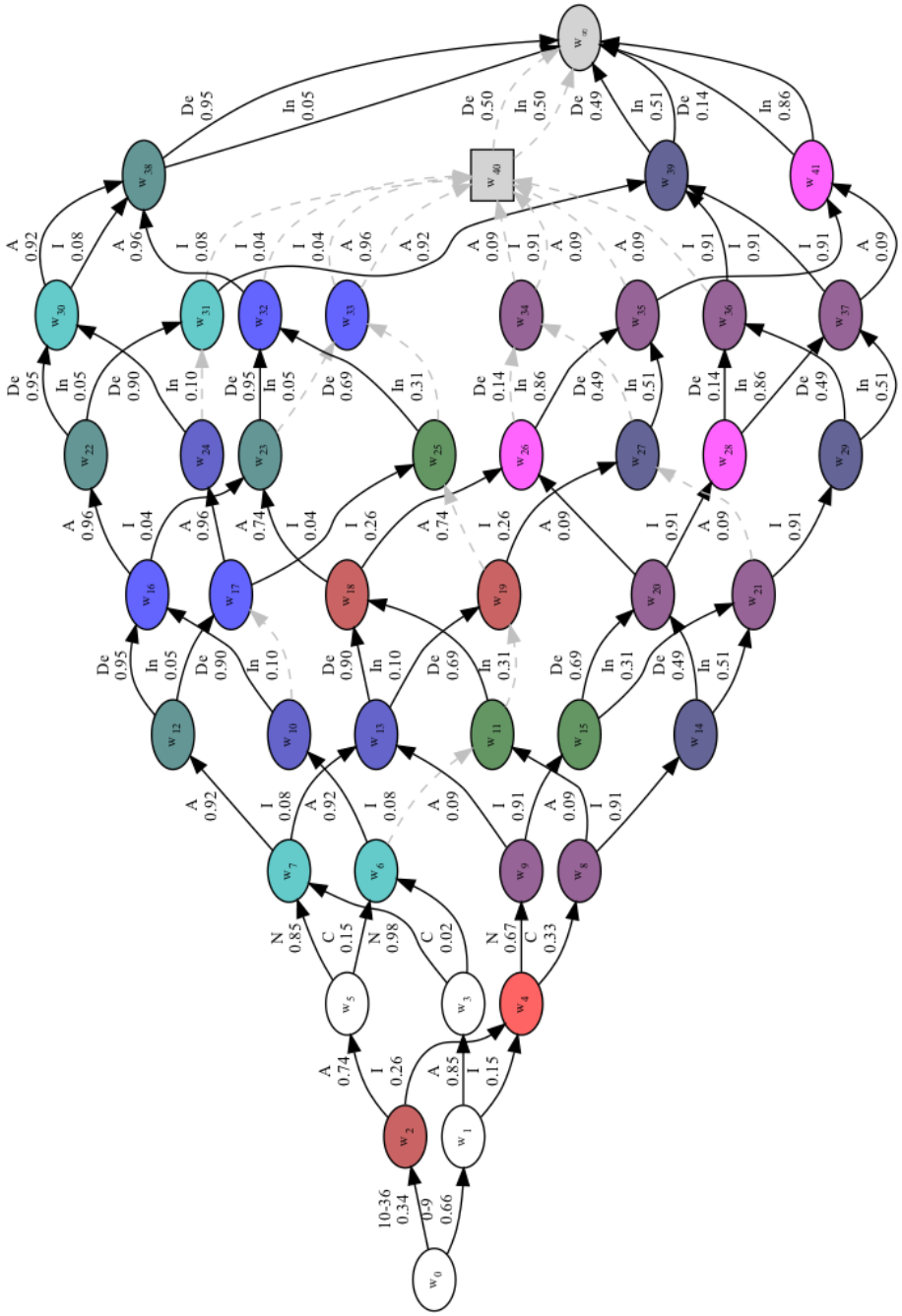


Fig. 12: Full CEG for the dentisty data with the Markov assumptions of Figure 11a. Edge labels are abbreviated - Adequate (A), Inadequate (I), Caries (C), No caries (N), Increase (In), Decrease (De).

Acknowledgements

Thank you to Gareth Walley and Aditi Shenvi for their valuable help in using and updating the `cegpy` package. Thank you to an anonymous referee for their helpful comments.

The research by JSC and ER was supported in part by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001 and the 100021-BIPE 2020 grant. ER acknowledges the financial support from the "Hub Life Science - Digital Health (LSH-DH) PNC-E3-2022-23683267 - Progetto DHEAL-COM ", which was granted by the Italian Ministry of Health as part of the Piano Nazionale Complementare for the "PNRR Ecosistema Innovativo della Salute".

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A Robustness experiment

In Section 3.5 we introduced an experiment to investigate the robustness of the full longitudinal ST and the longitudinal ST with Markov assumptions to small sample sizes. Here we give more details on this experiment and additional results.

The experiment was conducted using the depression data set which contains 340 samples. We begin by using the whole data set to select a full longitudinal ST model (as described in Section 3.1 and displayed in Figure 3), as well as a longitudinal ST with Markov assumptions (as described in Section 3.2 and displayed in Figure 5). We refer to these two selected models by ST-full and ST-Markov respectively. Associated to these models is a posterior distribution on the transition probabilities. The maximum a posteriori (MAP) estimates for the transition probabilities are also shown in the edge lables of these two figures.

Next we randomly choose 10% of the data to remove from the data set. For this smaller data set we fit a new full longitudinal ST model and longitudinal ST with Markov assumptions. We call these two selected models by ST-full-sub and ST-Markov-sub. Now we compare these selected models and their estimated transition probabilities to those fitted on the full data set. To compare the estimated transition probabilities we use the total variation distance (TVD) averaged over the 31 situations. For binary probabilities, the TVD is simply the difference between the two probabilities. To compare the selected models we consider the minimum number of changes in which one staging can be obtained from the other. For example, for the two stagings

$$\mathcal{S}_1 = ((s_1, s_2), (s_3, s_4), (s_5, s_6))$$

$$\mathcal{S}_2 = ((s_1, s_2, s_3), (s_4, s_5, s_6)$$

the distance is 2 because $\mathcal{S}_1$ can be obtained from $\mathcal{S}_2$ by changing the stage of s_3 and s_4. We call this the minimum change distance.

For a first experiment we compare both ST-full-sub and ST-Markov-sub to ST-full. This is because ST-full gives the best estimates of the individual transition probabilities. Hence this measures how well the two methods recover these best estimates when using a smaller sample size. For a second experiment we compare both methods to the same method on the whole data - that is comparing ST-full-sub to ST-full and ST-Markov-sub to ST-Markov. This is to see how sensitive each method is to reductions in sample size.

This was repeated 1000 times when removing 10% of the data, as well as another 100 times when removing 20%, ..., 90% of the data. Boxplots of the two distances for the first experiment setting are in Figure A1 and for the second experiment setting in Figure A2. In all cases the distance is higher for ST-Markov-sub than for ST-full-sub when only a small amount of data is removed. However, as more data is removed the distances associated to ST-full-sub quickly rise while those associated to ST-Markov-sub remain more stable. This is particularly the case for the estimated probabilities, but less so for the model selection. This shows that, for this depression data set, the addition of the Markov assumptions makes the probability estimates more robust when there is a small sample size.

References

Agresti, A. 2012. *Categorical data analysis*, Volume 792. John Wiley & Sons.

Ankinakatte, S. and D. Edwards. 2015. Modelling discrete longitudinal data using acyclic probabilistic finite automata. *Computational Statistics & Data Analysis* 88: 40–52 .

Barclay, L., R. Collazo, J. Smith, P. Thwaites, and A. Nicholson. 2015. The dynamic chain event graph. *Electronic Journal of Statistics 9*(2): 2130–2169 .

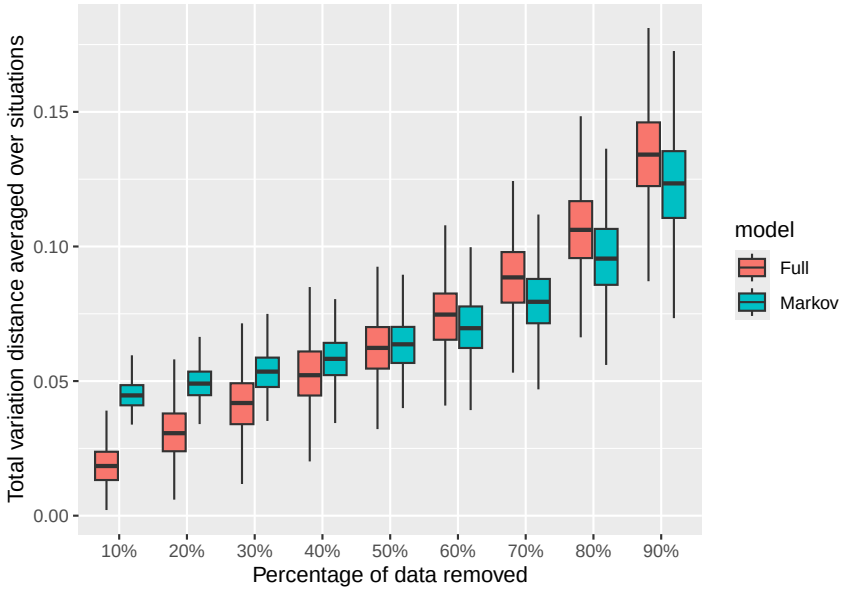
Barclay, L.M., J.L. Hutton, and J.Q. Smith. 2013. Refining a Bayesian network using a chain event graph. *International Journal of Approximate Reasoning 54*(9): 1300–1309 .

Bellazzi, R. and A. Riva. 1998. Learning Bayesian networks probabilities from longitudinal data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 28: 629–636 .

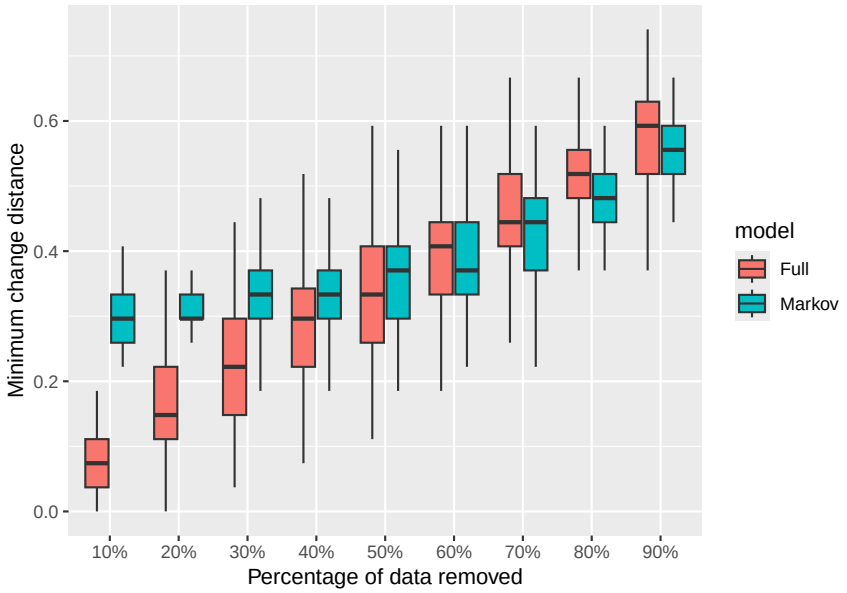
Carli, F., M. Leonelli, E. Riccomagno, and G. Varando. 2022. The R package stagedtrees for structural learning of stratified staged trees. *Journal of Statistical Software 102*(6): 1–30 .

Chen, R., S.M. Resnick, C. Davatzikos, and E.H. Herskovits. 2012. Dynamic Bayesian network modeling for longitudinal brain morphometry. *Neuroimage* 59: 2330–2338 .

Collazo, R.A., C. Görgen, and J.Q. Smith. 2018. *Chain event graphs*. CRC Press.

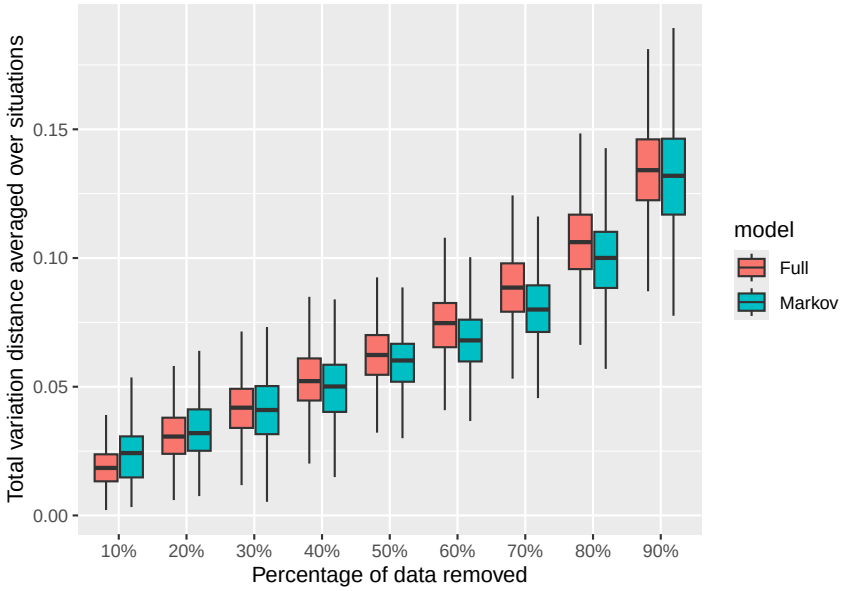


(a) Estimated probabilities

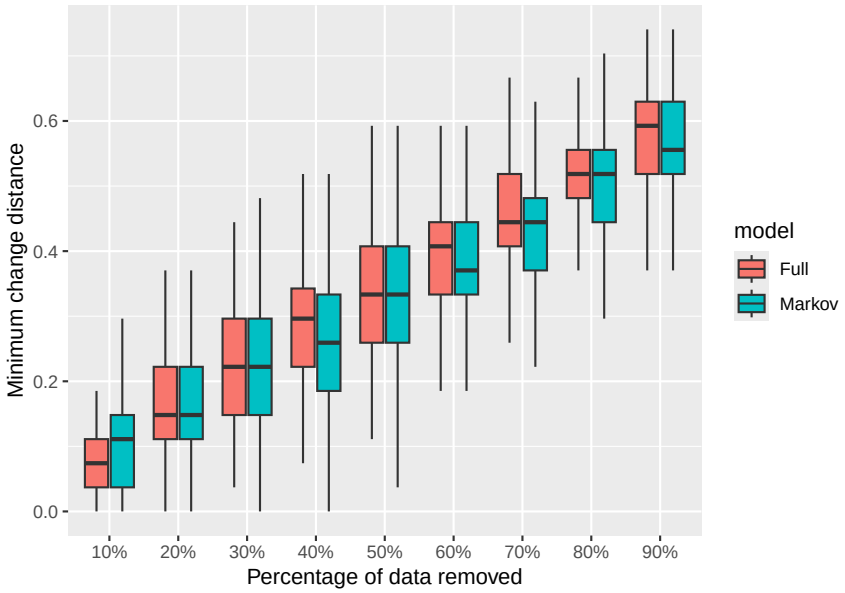


(b) Model selection

Fig. A1: Comparisons between ST-full-sub and ST-full (red) and between ST-Markov-sub and ST-full (blue).



(a) Estimated probabilities



(b) Model selection

Fig. A2: Comparisons between ST-full-sub and ST-full (red) and between ST-Markov-sub and ST-Markov (blue).

Collazo, R.A. and J.Q. Smith. 2016. A new family of non-local priors for chain event graph model selection. *Bayesian Analysis 11*(4): 1165–1201 .

Cowell, R. and J. Smith. 2014. Causal discovery through map selection of stratified chain event graphs. *Electronic Journal of Statistics 8*(1): 965–997 .

Diggle, P., P.J. Heagerty, K.Y. Liang, and S.L. Zeger. 2002. *Analysis of longitudinal data*. Oxford: Oxford University Press.

Freeman, G. and J. Smith. 2011a. Dynamic staged trees for discrete multivariate time series: forecasting, model selection and causal analysis. *Bayesian Analysis 6*(2): 279–306 .

Freeman, G. and J.Q. Smith. 2011b. Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis 102*(7): 1152–1165 .

Görgen, C., M. Leonelli, and O. Marigliano. 2022. The curved exponential family of a staged tree. *Electronic Journal of Statistics 16*(1): 2607–2620 .

Hernán, M.A., B. Brumback, and J.M. Robins. 2001. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association 96*(454): 440–448 .

Hughes, C., P. Strong, and A. Shenvi. 2022. Score equivalence for staged trees. *arXiv preprint arXiv:2206.15322* .

Hutton, J. 2015. Chain event graphs for missing data: exploring informative missing data, with examples from longitudinal studies. In *International Statistical Institute 60th World Statistics Congress*.

Jaeger, M., J.D. Nielsen, and T. Silander. 2006. Learning probabilistic decision graphs. *International Journal of Approximate Reasoning 42*(1-2): 84–100 .

Koch, G.G., J.R. Landis, J.L. Freeman, D.H. Freeman Jr, and R.G. Lehnen. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*: 133–158 .

Leonelli, M. and G. Varando 2022. Highly efficient structural learning of sparse staged trees. In *International Conference on Probabilistic Graphical Models*, pp. 193–204. PMLR.

Leonelli, M. and G. Varando. 2024. Structural learning of simple staged trees. *Data Mining and Knowledge Discovery*: 1–25 .

McGeachie, M.J., J.E. Sordillo, T. Gibson, G.W. Weinstock, Y.Y. Liu, D.R. Gold, S.T. Weiss, and A. Litonjua. 2016. Longitudinal prediction of the infant gut microbiome with dynamic Bayesian networks. *Scientific Reports* 6: 20359 .

Pensar, J., H. Nyman, T. Koski, and J. Corander. 2015. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data Mining and Knowledge Discovery* 29: 503–533 .

Prinzie, A. and D. Van den Poel. 2011. Modeling complex longitudinal consumer behavior with dynamic Bayesian networks: An acquisition pattern analysis application. *Journal of Intelligent Information Systems* 36: 283–304 .

Scutari, M. 2010. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software 35*(3): 1–22 .

Scutari, M., P. Auconi, G. Caldarelli, and L. Franchi. 2017. Bayesian networks analysis of malocclusion data. *Scientific reports 7*(1): 15236 .

Scutari, M. and J.B. Denis. 2021. *Bayesian networks: with examples in R*. CRC press.

Shenvi, A. and J.Q. Smith 2020. Constructing a chain event graph from a staged tree. In *International Conference on Probabilistic Graphical Models*, pp. 437–448. PMLR.

Silander, T. and T.Y. Leong 2013. A dynamic programming algorithm for learning chain event graphs. In *Proceedings of the 16th International Conference in Discovery Science*, pp. 201–216. Springer.

Smith, J.Q. and P.E. Anderson. 2008. Conditional independence and chain event graphs. *Artificial Intelligence 172*(1): 42–68 .

Strong, P. and J.Q. Smith 2022. Bayesian model averaging of chain event graphs for robust explanatory modelling. In *International Conference on Probabilistic Graphical Models*, pp. 61–72. PMLR.

Ugolini, A., F. Porro, F. Carli, P. Agostino, A. Silvestrini-Biavati, and E. Riccomagno. 2023, 10. Probabilistic graphical modelling of early childhood caries development. *PLOS ONE 18*(10): 1–14 .

Ugolini, A., S. Salamone, P. Agostino, E. Sardi, and A. Silvestrini-Biavati. 2018. Trends in early childhood caries: an italian perspective. *Oral Health Prev Dent 16*(1): 87–92 .

Varando, G., F. Carli, and M. Leonelli. 2024. Staged trees and asymmetry-labeled DAGs. *Metrika*: 1–28 .

Walley, G., A. Shenvi, P. Strong, and K. Kobalczyk. 2023. cegpy: Modelling with chain event graphs in Python. *Knowledge-Based Systems* 274: 110615 .

Zeger, S.L. and K.Y. Liang. 1992. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine 11*(14-15): 1825–1839 .