On Bayesian Dirichlet Scores for Staged Trees and Chain Event Graphs

Conor Hughes 1
Peter Strong 2,3
Aditi Shenvi 1

1 Statistics Department, University of Warwick, Coventry, United Kingdom
2 Centre for Complexity Science, University of Warwick, Coventry, United Kingdom
3 The Alan Turing Institute, London, United Kingdom

Abstract

Chain event graphs (CEGs) are a recent family of probabilistic graphical models that have emerged as a suitable alternative to Bayesian networks (BNs) for asymmetric processes. These asymmetries include context-specific independencies and structural asymmetries (i.e. structural zeros and structural missing values). Model selection in CEGs is done through an intermediate model called staged trees, and similar to BNs, this can be done through a score-based approach. Moreover, a CEG is uniquely defined by its staged tree. In BNs, the Bayesian Dirichlet equivalent uniform (BDeu) score – obtained through a specific hyperparameter setting in the Bayesian Dirichlet score function – is popular for score-based model selection for its desirable theoretical properties such as ease of hyperparameter setting, preservation of effective sample size, and score equivalence. It has been shown that, under standard assumptions, the BD score function can analogously be defined for staged trees and thereby, for CEGs. However, unlike in BNs, there has been little research into the effects of hyperparameter setting in the BD score function on both these models. In this paper, we derive a BDeu score for staged trees and CEGs. Further, we explore the relationship between the BD sparse (BDs) score, proposed for BNs that contain unobserved configurations of its variables within a dataset, and the BDeu for staged trees and CEGs. Through this relationship, we demonstrate the favourable properties of CEGs in modelling processes with sparsity or asymmetry.

Keywords: chain event graphs; event trees; model selection; structural zeroes; structural missing values; directed graphical models.

1. Introduction

Chain event graphs (CEGs) are probabilistic graphical models that can represent complex independence statements between variables on discrete data. They contain the popular graphical modelling family of discrete Bayesian networks (BN) as a subclass (Barclay et al., 2013). Unlike BNs, CEGs are able to model asymmetric processes. We say that a process is asymmetric when it exhibits at least one of two types of asymmetries. The first is asymmetric conditional independence relationships or context-specific conditional independencies which are independencies of the form $X \perp Y | Z = z_1$ but $X \not\perp Y | Z = z_2$ for some variables $X$, $Y$ and $Z$ where $\perp$ stands for probabilistic independence and the vertical bar shows conditioning variables on the right. These are expressed in the graph of a CEG through the colouring of its nodes. The second is structural asymmetry which occurs when the event space of a process does not conform to a product space structure (Shenvi, 2021). Structural asymmetries occur naturally or by design in many domains, such as public health (Shenvi and Smith, 2021).
Hughes, Strong and Shenvi (2020), security (Shenvi et al., 2020) and migration (Strong et al., 2021). CEGs can accommodate these through their tree-based constructions.

CEGs are obtained from event trees, through the construction of an intermediate model called staged trees (Smith and Anderson, 2008). More specifically, model selection for CEGs is equivalent to obtaining the staged tree from the event tree as the mapping from the staged tree to the CEG is bijective (Shenvi and Smith, 2020; Smith and Anderson, 2008). Hereafter, for simplicity, we will focus on staged trees. Note, however, that the advances in this paper are equally applicable to CEGs.

Event trees can be a more natural way of depicting a process as an unfolding of events (Shafer, 1996). Staged trees are created from event trees by embellishing their situations, i.e. non-leaf nodes, with colour. Situations are assigned the same colour if they have the same conditional transition probability and we define situations assigned the same colour as being in the same stage.

Example 1 Suppose there is a study on the length of hospital stays for women after giving birth. The response variable is $X_L$, a binary variable recording whether an individual was in hospital for $\leq 2$ days (value 1) or $> 2$ days (value 0). Three covariates were also recorded:

- $X_A$: The age of the individual: either $\geq 29$ years (1) or $< 29$ years (0);
- $X_P$: The length of the pregnancy term: either $\geq 37$ weeks (1) or $< 37$ weeks (0);
- $X_B$: The type of birth: either without (1) or with (0) a caesarean section.

Figure 1 shows the directed acyclic graph (DAG) of the BN for this study based on the following independencies:

$$X_A \perp \perp X_P$$
$$X_L \perp \perp X_P \mid X_A, X_B$$

Note that if we were to assign a strict ordering on the variables of the DAG, we could equivalently place $X_P$ or $X_A$ first. Now, suppose that recent studies show that the length of hospital stay was independent of age only when a c-section didn’t occur, but not when it did. This can be written as

$$X_L \perp \perp X_A \mid X_B = 1.$$  \hspace{1cm} (1)

This is an example of a context-specific conditional independence. The DAG in Figure 1 cannot represent this information without serious modifications.

We demonstrate now that through the use of colour, a staged tree can encode this information in its graph. Figure 2 gives two staged trees – $T_P$ with $X_P$ at the root and $T_A$ with $X_A$ at the root – both of which can represent Statement 1. Both of these trees are statistically equivalent (Görgen and Smith, 2018). Note that stages with only one situation have their colouring suppressed.
In both staged trees, edges with a corresponding label of 1 are directed upwards, and the edges in the tree have been labelled with the counts of individuals who traverse each edge. For example, the edge from $s_3$ to $s_7$ in $T_P$ represents that, in the sample, there are 10 women under the age of 29 with a term over 37 weeks and no c-section.

Next, suppose that for any older individual with a pregnancy term less than 37 weeks (the situation $s_6$), the birth must necessarily be a c-section. This would imply that a birth without a c-section in such a scenario would be a structural zero (i.e., a logical impossibility), and so the count on the edge from $s_6$ to $s_{13}$ will always be zero: any individual at $s_6$ must also be at $s_{14}$. Furthermore, suppose an additional variable recording whether a miscarriage occurred, $X_M$, was introduced. Then, for women who suffer a miscarriage, $X_B$ is meaningless. This constitutes a structural missing value which is represented within the event tree by excluding any edges for $X_B$ for women who suffer a miscarriage. These are both examples of structural asymmetries, which cannot be represented by a BN but can be by a staged tree.

Various model selection methodologies have been explored for staged trees. Often, these methodologies are extensions of the structure learning techniques for BNs, which can be score-based,
Hughes, Strong and Shenvi

constraint-based or hybrid (Scutari, 2018). Constraint-based methods use statistical tests to learn conditional independence statements. Score-based methods are the more classical model selection methods aiming to maximise some objective function, such as the marginal likelihood or non-Bayesian methods including the Bayesian Information Criterion (Schwarz, 1978) or Akaike Information Criterion (Akaike, 1974). Hybrid methods are a combination of the two approaches.

For staged trees, score-based methods are the most common thus far, see Carli et al. (2022) for an example of hybrid model selection that fits a BN first, but very little research has been done on the properties of the scoring functions used. Score-based methods for BNs often have desirable properties for the score functions, such as ease of setting prior hyperparameters (particularly when no domain expertise is available), and score equivalence for equivalent probabilistic structures (Heckerman et al., 1995).

In BNs under certain assumptions the Bayesian Dirichlet score equivalence (BDe) metric (Heckerman et al., 1995) is score equivalent and when an effective sample size $\alpha$, a measure of the strength of one’s prior belief, is propagated uniformly through the tree, we get the Bayesian Dirichlet equivalent uniform (BDeu) score. This only requires the setting of one hyperparameter, and still satisfies score equivalence (Chickering, 1995), while essentially being a weakly informative prior. For the BDeu, this effective sample size is ideally kept consistent for each random variable. However, when values of some $X_i$ are unobserved for certain parent combinations, the effective sample size decreases, leading to overfitting and the inclusion of spurious arcs (Scutari, 2016). This can be a problem with sparse data, where the sample size is not large compared to the number of random variables and parent combinations. Scutari (2016) proposes an intuitive adjustment called the Bayesian Dirichlet sparse (BDs) score, where only non-zero parent combinations are considered and the effective sample size is maintained throughout. This has several nice properties for model selection, but is only asymptotically score equivalent (Scutari, 2018).

The purpose of this paper is two-fold. First, we extend the BDeu to staged trees, such that the desirable properties are maintained whilst also being applicable to a wider class of staged trees than current formulations. Second, we extend the BDs approach to staged trees using the explicit treatment of zeroes alongside this newly formulated BDeu. In Section 2 we introduce notation for BNs and staged trees. In Section 3 we review the BDeu and the BDs for BNs. In Section 4 we propose a new formulation of BDeu for staged trees. In Section 5 we discuss the relation between BDs and staged trees. We conclude with a discussion in Section 6.

2. Notation and Preliminaries

2.1 Bayesian Networks

Let $D$ denote the complete random data sample used. We define a BN $\mathcal{B} = (\mathcal{G}, \pi)$ over the random variables $X = (X_1, \ldots, X_n)$ as a DAG $\mathcal{G}$ whose nodes are given by the variables $X$ and whose edge set consists of directed arcs such that lack of an edge between two nodes represents conditional independence between the variables represented by the nodes. Further, the dependencies in the BN are quantified by the conditional probability tables $\pi$. For a variable $X_i$, let $r_i$ denote the size of its state space, $q_i$ denote the size of its parent state space, and let the Dirichlet hyperparameter for its $j^{th}$ parent combination be $\alpha_{ij} = (\alpha_{ij1}, \ldots, \alpha_{ijr_i})$. Let $\alpha_{ij} = \sum_k \alpha_{ijk}$. Let the corresponding observed number of individuals for variable $X_i$, parent combination $j$ and state $k$ be $n_{ijk}$, with vector of counts $n_{ij} = (n_{ij1}, \ldots, n_{ijr_i})$ and $\overline{n}_{ij} = \sum_k n_{ijk}$. For a graph $\mathcal{G}$ and complete sample $D$, suppose that the positivity, parameter independence and parameter modularity assumptions hold. Then, with prior
conditional transition probability vectors. The collection of stages partitions the node set of the expectation of a Dirichlet random variable, which is the posterior distribution (Heckerman et al., 1995), decreases the variance of the posterior parameter estimates due to it appearing in the denominator for edges, from the root. A staged tree is said to be stratified of conditional transitional probability vectors for its nodes. A \( X \) of each \( f \) a di product space of a pre-specified collection of variables. Here, each level of the tree corresponds to a different variable, and each situation at that level has the same number of outgoing edges, one for each state. We define a staged tree as any staged tree which is not stratified.

\[ \alpha \]

\[ \theta \]

2.2 Staged Trees

We define a staged tree as \( (T, \theta_T) \) where \( T \) is the underlying event tree, and \( \theta_T \) is the collection of conditional transitional probability vectors for its nodes. A leaf is any node with no outgoing edges and a situation is any non-leaf node. Define a stage \( u \) as a set of situations with equivalent conditional transition probability vectors. The collection of stages partitions the node set of \( T \).

A level of a staged tree is the collection of situations which are the same distance, or number of edges, from the root. A staged tree is said to be stratified when the event space of the process is the product space of a pre-specified collection of variables. Here, each level of the tree corresponds to a different variable, and each situation at that level has the same number of outgoing edges, one for each state. We define a non-stratified staged tree as any staged tree which is not stratified.

Let \( \Lambda(T) \) be the set of all root-to-leaf paths in tree \( T \). For a path \( \lambda \in \Lambda(T) \), let \( E(\lambda) \) be the edge set of that path. An edge \( e \) can also be written as the pair \((v_m, v_n)\), where \( e \) is the edge between the nodes \( v_m \) and \( v_n \). For simplicity and clarity, we can write this edge as \( e_{m,n} \). Define the situation-centred event at situation \( v \) as \( \Lambda(v) = \{ \lambda \in \Lambda(T) : \exists (v, \cdot) \in E(\lambda) \} \). For an edge \( e \), define the edge-centred event as \( \Lambda(e) = \{ \lambda \in \Lambda(T) : \exists e \in E(\lambda) \} \).

3. BDeu and BDs for BNs

3.1 Review

The BD score in Equation \ref{eq:bd} is dependent on the structure of the BN, namely \( \text{pa}(i) \), the parent set of each \( X_i \), and the choice of the hyperparameters \( \alpha_{ijk} \). For some choice of \( \alpha \), variously called the effective, imaginary or equivalent sample size, let \( \alpha_{ijk} = \alpha/(r_{ij}) \). This choice of hyperparameter gives the BDeu metric (Heckerman et al., 1995). The effective sample size for each variable \( X_i \) is \( \overline{\alpha_i} = \sum_{ijk} \alpha_{ijk} = \alpha \), and so is invariant with \( i \). In addition, the BDeu metric is score equivalent which implies that the BDeu would give the same score to two DAGs that encode the same probability distribution (Chickering, 1995). This is a particularly desirable result, as often causal relationships (and thus strict ordering of arcs and variables) are not explicitly known, and so there can be several DAGs that are in the same equivalence class, and that should not be discriminated from each other. Furthermore, the BDeu metric often is used alongside a Uniform prior on the model space, where every model is given equal prior probability (Heckerman et al., 1995).

The BDeu metric only requires one hyperparameter specification, \( \alpha \), and so is easily generated without much prior information. However, it is not without its limitations (see Scutari (2016)). This includes the fact that the sensitivity to choices of \( \alpha \) results in a prior that is not as weakly informative as one would like, particularly for medium-sized data (Steck and Jaakkola, 2003). Furthermore, it is common to associate the choice of \( \alpha \) with the strength of one’s prior beliefs, and increasing \( \alpha \) decreases the variance of the posterior parameter estimates due to it appearing in the denominator for the expectation of a Dirichlet random variable, which is the posterior distribution (Heckerman et al., 1995).
3.2 Consistency Issues

Consider the case where \( \pi_{ij} = 0 \). In the BDs setting, each outgoing edge will have a hyperparameter \( \alpha_{ijk} = 0 \). This is equivalent to treating it as a structural zero. However, \( \pi_{ij} = 0 \) means that at least one of the states of one of its parents in this \( j \textsuperscript{th} \) combination is unobserved. This unobserved parent combination may be a sampling or structural zero. As the first variable in the BN has no parents, it will always have nonzero hyperparameters. Using this and the fact that any zero hyperparameter must follow a zero count for one of its parents, when \( \pi_{ij} = 0 \), there will always be a zero count with a non-zero prior hyperparameter amongst its ancestors. This could be thought of as the first zero in the chain of the process. This initial zero, with its non-zero hyperparameter, is treated as a sampling zero, but its descendants, which will be zero as a result, are treated as structural zeroes.

We believe this inconsistency is why the BDs is not score equivalent, as for models with different variable orderings, a different zero count will be treated as a sampling zero even if these competing models are supposed to be in the same equivalence class.

Example 1 (continued) We refer to the DAG in Figure 1. Suppose there were no individuals over 29 years of age observed, but there were younger individuals. In this case, \( \alpha_A = \alpha, \alpha_{A,1} = \alpha_{A,0} = \alpha/2 \), and so the older individuals are a sampling zero. Necessarily, because there were no individuals over 29 observed, there were also no births observed for those over 29, and so two of the four parent combinations for \( X_B \) are unobserved. Thus, the hyperparameters for these parent combinations will be zero, they will be assumed to be structural zeroes, and they will not contribute to the
BDs metric. However, no such births were observed because no older individuals were observed, yet the reasons for their absences are being considered differently.

Refer now to \( T_P \) in Figure 2 and assume no individuals over 29 are observed, so the edges \( e_{1,4} \) (the edge from situation \( s_1 \) to \( s_4 \)) and \( e_{2,6} \) have zero counts. Calculating the hyperparameters in the same way as for the BDs, these edges will have nonzero hyperparameters, corresponding to sampling zeroes, but the situations they enter, \( s_4 \) and \( s_6 \), will not be observed. Their outgoing edges, say \( e_{4,9} \), will now have \( \alpha_{4,9} = 0 \), and they will be a structural zero. Labelling the edges of the tree with the hyperparameters, and picturing the hyperparameters propagating along the edges of the tree, some edges go from nonzero to zero, and others increase to account for them in maintaining the effective sample size. Hence, the issues with consistency in definitions of zeroes can be seen clearly through a staged tree.

4. BDeu for Staged Trees

The agglomerative hierarchical clustering (AHC) algorithm is the most common score-based method to merge situations into stages at each level of the tree (Freeman and Smith, 2011). The BD metric itself is a popular choice of scoring function (Freeman and Smith, 2011), and Cowell and Smith (2014) proposed a BDeu analogue for CEGs. However, this BDeu analogue was notationally lacking, and did not have the flexibility to cover the entire staged tree family, which is larger than the BN family. In particular, it is only applicable to stratified staged trees, which has a direct, although sometimes approximate, translation as a BN. For example, in a staged tree, a structural zero can be represented by deleting the associated edge (Shenvi, 2021), and so the underlying event tree is no longer stratified. The BDeu analogue to this tree would result in overestimation of the number of edges in the tree and subsequently the effective sample size would not be maintained.

We propose a new way of setting the hyperparameters for the BDeu for CEGs which is centred on path uniformity, where every possible path is equally probable. We assign each path a uniform prior weight based on the effective sample size and number of paths, and back-propagate this through the staged tree to obtain the prior hyperparameters for each situation and its edges, \( \alpha_{ij} \) and \( \alpha_{ijk} \) respectively. This is motivated by the uniform prior on paths mentioned in Barclay et al. (2013).

There is a one-to-one correspondence between the root-to-leaf paths in a tree, and its leaf nodes, as each leaf represents a unique unfolding of the process, so these can also be referred to interchangeably. Let \( L \) be the number of leaf nodes; that is \( L = \#\Lambda(T) \). For each leaf node, we assign a weight of \( \frac{\alpha}{L} \). Then, for each situation and edge in the tree, the prior hyperparameter is simply the sum of the prior weights of the leaf nodes descendant from it. With uniform weights, this is simply the product of \( \frac{\alpha}{L} \) and the size of situation- or edge-centred event. Consider a situation \( s_{ij} \), the \( j^{th} \) situation at level \( i \), with \( r_{ij} \) outgoing edges. Then, the hyperparameter for its \( k^{th} \) outgoing edge \( e_{ijk} \) is

\[
\alpha_{ijk} = \frac{\alpha}{L} \#\Lambda(e_{ijk}).
\]

The hyperparameter for the situation \( s_{ij} \) can be defined similarly. Using the fact that:

\[
\#\Lambda(s_{ij}) = \sum_{k=1}^{r_{ij}} \#\Lambda(e_{ijk}),
\]

1. The translation will be approximate in the presence of context-specific information.
we also have that $\alpha_{ij} = \sum_k \alpha_{ijk}$ as desired.

Now, this BDeu metric will be calculated stage-wise, not situation-wise. That is, rather than summing over the contributions for each situation at a level, the contributions for each stage are considered instead. Suppose a stage $u_j$ consists of $N_j$ situations, labelled $s_j^{(l)}$, $l = 1, \ldots, N_j$. Then, the edge hyperparameters for $u_j$ are

$$\alpha_{jk} = \frac{\alpha}{L} \sum_{l=1}^{N_j} \# \Lambda(e_{jk}^{(l)}). \tag{3}$$

The hyperparameter for the stage itself is $\overline{\alpha_j} = \sum_{l=1}^{N_j} \alpha_{jk}$. This extends similarly to the counts $n_j$ for the stage. Then, in a staged tree where all leaves are the same distance from the root, and where no stage is on multiple levels, for the level $i$ which has $J_i$ stages, we have that

$$\overline{\alpha_i} = \sum_{j=1}^{J_i} \sum_{k=1}^{r_j} \alpha_{jk} = L \frac{\alpha}{L} = \alpha.$$

Note that the effective sample size is maintained for the applicable staged trees described above, which include all stratified staged trees. When there is an asymmetric process with paths of different lengths, each time a path is concluded, the effective sample size for the tree as a whole will decrease by the weight attributed to this path. This could be accounted for by, at each level, weighting each path based on the remaining number of leaves rather than total number of leaves in the tree, in a way an analogue to the BDs. However, doing this violates the conservation of the hyperparameters moving from edge to situation that is necessary for score equivalence, as discussed in Section 6. Furthermore, if a stage is on multiple levels, which is possible for a non-stratified staged tree, then interpreting the BDeu by level can lead to overestimation of the effective sample size. As such, it is better to look at the collection of stages as a whole, rather than by level.

Once the hyperparameters have been set, the BDeu metric for a staged tree $T$ with $J$ stages (the most general form) can now be formulated as:

$$\text{BD}(T, D; \alpha) = \prod_{j=1}^{J} \left[ \frac{\Gamma(\overline{\alpha_j})}{\Gamma(\overline{\alpha_j} + n_j)} \prod_{k=1}^{r_j} \frac{\Gamma(\alpha_{jk} + n_{jk})}{\Gamma(\alpha_{jk})} \right]. \tag{4}$$

**Example 1 (continued)** Revisiting the staged tree $T_p$ for the pregnancy example in Figure 2, there are 16 leaf nodes, corresponding to $16 = 2^4$ root-to-leaf paths. For the edge between $s_3$ and $s_7$, so an individual under 29 with a term over 37 weeks and no c-section (which can be associated with $s_7$ itself), there are 2 leaves that are descendant, and so $\# \Lambda(s_7) = 2$, $\alpha_{3,7} = \overline{\alpha_7} = \frac{2\alpha}{16} = \frac{\alpha}{8}$. This should not be surprising considering there are 4 singleton stages, so 8 edges in total to divide the effective sample size across.

However, now look at the stage $u_{red} = \{s_1, s_2\}$. Each edge has 4 descendant leaves, but looking at the edges for the stage as a whole, each of the two edges has $4 + 4 = 8$ descendant leaves, and so the prior hyperparameter of each edge is $\frac{\alpha}{2}$. This level only contains one stage and so the effective sample size is still $\alpha$.

Recall the equivalent depiction $T_A$ in Figure 2. In both trees, the levels for $X_B$ and $X_L$ are the same, except possibly reordered vertically, and their product will be the same. We focus on
calculating the BDeu metric for just the first two levels, for a general $\alpha$. For $T_P$, this is:

$$B_{\text{Deu}}(X_P, X_A) = \left[ \frac{\Gamma(\alpha)}{\Gamma(\alpha + 40)} \frac{\Gamma(\alpha/2 + 25)\Gamma(\alpha/2 + 15)}{\Gamma(\alpha/2)^2} \right] \left[ \frac{\Gamma(\alpha)}{\Gamma(\alpha + 40)} \frac{\Gamma(\alpha/2 + 24)\Gamma(\alpha/2 + 16)}{\Gamma(\alpha/2)^2} \right]$$

(5)

Meanwhile, for $T_A$ this is:

$$B_{\text{Deu}}(X_A, X_P) = \left[ \frac{\Gamma(\alpha)}{\Gamma(\alpha + 40)} \frac{\Gamma(\alpha/2 + 24)\Gamma(\alpha/2 + 16)}{\Gamma(\alpha/2)^2} \right] \left[ \frac{\Gamma(\alpha)}{\Gamma(\alpha + 40)} \frac{\Gamma(\alpha/2 + 25)\Gamma(\alpha/2 + 15)}{\Gamma(\alpha/2)^2} \right]$$

(6)

which is the exact same as (5), and the two statistically equivalent trees have the same score.

This new formulation of the BDeu metric is essentially a different way of expressing the BDeu initially proposed in Cowell and Smith (2014), but now versatile enough to be applied to non-stratified staged trees. This flexibility will be advantageous when addressing zero counts.

5. BDs for Staged Trees

We now attempt to extend the BDs to staged trees. Unlike BNs, staged trees require the user to be explicit with regards to their treatment of zero counts as sampling zeros or structural zeros. The latter is represented by deleting the corresponding edge in the tree. As the tree depicts the entire unfolding of the process, any inconsistencies will raise questions on the logic behind the underlying prior assumptions.

For staged trees, a differential treatment of a particular zero count on a path between competing models will necessarily take us out of the equivalence class and therefore, will generally result in different scores for the models. To stay within the equivalence class of a staged tree, all competing models must treat each zero count consistently; either as a structural zero, or as a sampling zero. This could be interpreted as an assertion on the path itself. Due to the versatility of the staged tree, these decisions need not be uniform across all paths, i.e. one zero can be structural whilst another can be sampling. However, in the absence of any prior information, or a known logical unfolding of the process, we argue that treating all zeros in the same manner is beneficial.

Suppose we characterise each zero count in the staged tree as a sampling zero. This is in fact the usual BDeu metric without any adjustments, and thus runs into the same problems as for the BN. Similarly, defining any single path as a sampling zero will still decrease the effective sample size once the first zero on the path occurs. In order to maintain the effective sample size whilst also being consistent, we could instead define each of the zero counts as a structural zero. We then define an adjusted tree $T^*$ as $T$ when each edge that is assumed to be a structural zero is deleted. Necessarily, this enforces a certain stage structure in the tree, as situations can only be in the same stage as other situations which have the same number of edges, and same edge labels. Once this stage structure is enforced, any staged tree which includes this stage structure can now be compared.

This methodology demonstrates part of the motivation for the expression of the BDeu in terms of leaves. The adjusted tree $T^*$ is a non-stratified staged tree once any edge is deleted unless all identical edges for a variable are deleted, which would raise questions on the variable’s inclusion in the first place. Hence, we now define the staged tree equivalent of the BDs as the BD structurally sparse (BDstruc) metric, which is simply the BDeu applied to the adjusted tree. The effective sample size will be maintained (in equivalent scenarios as for the BDeu), one of the incentives for the BDs. In addition, if the BDeu is score equivalent, so is the BDstruc, due to the consistency in
treatment of zeroes across competing models within the equivalence class. We thus propose that when no prior information is available, the BDstruc should be used.

**Example 1 (continued)** We refer again to the staged tree $T_P$ in Figure 2 and assume each zero count is a structural zero. This deletes the edge from $s_6$ to $s_{13}$, which also deletes $s_{13}$, $l_{13}$ and $l_{14}$. We also delete the edge from $s_{11}$ to $l_{10}$, which deletes $l_{10}$, and removes $s_{11}$ from the same stage as $s_7$ and $s_9$. The adjusted staged tree, $T_P^*$ can be seen in Figure 3. Note that the indexing for the situations and leaves has been updated to account for the deletions. Furthermore, $s_9$ is in its own stage. Now, the levels for $X_B$ and $X_L$ no longer have the same number of edges for each node, and the adjusted tree is no longer stratified. The BDstruc is now the BDeu applied to this non-stratified tree with all structural zeroes. In this case, $L = 13$, and $\alpha_3 = \frac{4}{13}$, while $\alpha_5 = \frac{3}{13}$ and $\alpha_6 = \frac{2\alpha}{13}$. The hyperparameters still sum to $\alpha$ as desired, but now the situations $s_5$ and $s_6$ have a lower prior belief than $s_3$ and $s_4$. Intuitively, this makes sense; we are assuming each path has an equal probability of traversal and so situations on less paths have a lower prior probability of being visited. Consequently, the situations and edges that share paths with $s_5$ and/or $s_6$ will also have lower prior beliefs. In particular, for the edge from $s_0$ to $s_1$ we have that $\alpha_{0,1} = \frac{8\alpha}{13}$, but $\alpha_{0,2} = \frac{5\alpha}{13}$. The BDstruc score can be calculated, and when compared to $T_A$ (through the adjusted tree $T_A^*$) in a similar way to the BDeu, they have the same score.

![Figure 3: The adjusted staged tree $T_P^*$ for the pregnancy example assuming all zeroes are structural zeroes. The node indices have been updated.](image-url)
6. Discussion

In this paper we proposed a new formulation of the hyperparameters for the BDeu metric for staged trees and CEGs. This metric is now capable of being applied to non-stratified staged trees, which often occur in applications of real-world systems. This greater applicability can be combined with the assumptions of structural zeroes and the creation of an adjusted staged tree to calculate the BDstruc metric. The BDstruc metric, as with the BDs, maintains the effective sample size throughout for staged trees where each path is of the same length and stages are not on multiple levels.

We posit that the BDeu is score equivalent, and the next step is to publish a formal proof. The statistical equivalence classes of staged trees can be traversed through two operators, the swap and resize, outlined in Görgen and Smith (2018) and Görgen et al. (2022). The swap operator reorders suitable subtrees, and the BDeu metric on these subtrees is invariant by commutativity. Examples of suitable subtrees, called twins, can be seen in Figure 2 for the subtrees containing $X_P$ and $X_A$, where their order is reversed. The score equivalence of these two trees was shown in Section 4. Furthermore, the resize operator transforms suitable subgraphs into subtrees where each root-to-leaf path is one edge, or vice versa. These suitable subgraphs come in two cases but are in essence saturated subgraphs, across which the BDeu simplifies nicely.

In staged trees, issues with sparsity are at their worst for the final level, the edges that enter the leaves. In the BDstruc, we have proposed treating these zeroes as structural zeroes, in the same way as for all other edges. However, when an edge entering a leaf has a zero count, but the situation it is emanating from has a non-zero count, there are arguments for both treatments. Treating it as a structural zero and deleting the edge is the simplest way, consistent with the rest of the BDstruc, and hence maintains any score equivalence should this level be reordered. It also avoids any spurious merging with other stages. However, often this final level is logically placed, either as a designated response variable or obvious end of the process, and so reordering is not considered. Moreover, the effective sample size will not be affected, as it is the final level. In addition, while spurious merging is avoided, so is all merging, and valuable stage information could be lost, particularly if it was a sampling zero in reality. This is a common occurrence when the tree is sparse. Finally, in a sparse tree, many edges may be deleted if they were designated as structural zeroes, which can significantly affect the prior hyperparameters throughout the tree, due to back-propagation, and also limit interesting stage discoveries for the final level. As such, whilst we propose the BDstruc as a generally applicable solution, we suggest strong consideration should be given to the treatment of each zero count in particular.

Acknowledgments

The authors would like to thank Dr Christiane Görgen and Prof Jim Q. Smith for discussions and suggestions that have greatly helped to improve this paper.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
L. M. Barclay, J. L. Hutton, and J. Q. Smith. Refining a Bayesian network using a chain event graph. *International Journal of Approximate Reasoning*, 54(9):1300–1309, 2013.

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

D. M. Chickering. A transformational characterization of equivalent bayesian network structures. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 87–98, 1995.

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

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

C. Görgen and J. Q. Smith. Equivalence classes of staged trees. *Bernoulli*, 24(4A):2676–2692, 2018.

C. Görgen, A. Maraj, and L. Nicklasson. Staged tree models with toric structure. *Journal of Symbolic Computation*, 113:242–268, 2022.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

M. Scutari. An empirical-bayes score for discrete bayesian networks. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models*, pages 438–448. PMLR, 2016.

M. Scutari. Dirichlet bayesian network scores and the maximum relative entropy principle. *Behaviorometrika*, 45(2):337–362, 2018.

G. Shafer. *The art of causal conjecture*. MIT Press, 1996.

A. Shenvi. *Non-Stratified Chain Event Graphs: Dynamic Variants, Inference and Applications*. PhD thesis, University of Warwick, 2021.

A. Shenvi and J. Q. Smith. Constructing a chain event graph from a staged tree. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models*. PMLR, 2020.

A. Shenvi, F. O. Bunnin, and J. Q. Smith. A bayesian decision support system for counteracting activities of terrorist groups. arXiv preprint arXiv:2007.04410, 2020.

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

H. Steck and T. Jaakkola. On the dirichlet prior and bayesian regularization. *Advances in Neural Information Processing Systems*, 15:713–720, 2003.

P. Strong, A. McAlpine, and J. Q. Smith. A bayesian analysis of migration pathways using chain event graphs of agent based models. arXiv preprint arXiv:2111.04368, 2021.