Journal of Biomedical Informatics 61 (2016) 283–297

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Understanding disease processes by partitioned dynamic Bayesian networks

Marcos L.P. Bueno [a,*], Arjen Hommersom [a,c], Peter J.F. Lucas [a,b], Martijn Lappenschaar [a], Joost G.E. Janzing [d]

[a] Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands
[b] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
[c] Faculty of Management, Science and Technology, Open University, The Netherlands
[d] Department of Psychiatry, Radboud University Nijmegen Medical Center, The Netherlands

ARTICLE INFO

Article history:
Received 19 November 2015
Revised 4 April 2016
Accepted 11 May 2016
Available online 12 May 2016

Keywords:
Probabilistic graphical models
Dynamic Bayesian networks
Non-homogeneous stochastic processes
Heuristic algorithm
Multivariate time series
Psychotic depression

ABSTRACT

For many clinical problems in patients the underlying pathophysiological process changes in the course of time as a result of medical interventions. In model building for such problems, the typical scarcity of data in a clinical setting has been often compensated by utilizing time homogeneous models, such as dynamic Bayesian networks. As a consequence, the specificities of the underlying process are lost in the obtained models. In the current work, we propose the new concept of partitioned dynamic Bayesian networks to capture distribution regime changes, i.e. time non-homogeneity, benefiting from an intuitive and compact representation with the solid theoretical foundation of Bayesian network models. In order to balance specificity and simplicity in real-world scenarios, we propose a heuristic algorithm to search and learn these non-homogeneous models taking into account a preference for less complex models. An extensive set of experiments were ran, in which simulating experiments show that the heuristic algorithm was capable of constructing well-suited solutions, in terms of goodness of fit and statistical distance to the original distributions, in consonance with the underlying processes that generated data, whether it was homogeneous or non-homogeneous. Finally, a study case on psychotic depression was conducted using non-homogeneous models learned by the heuristic, leading to insightful answers for clinically relevant questions concerning the dynamics of this mental disorder.

1. Introduction

Understanding the evolution of disease processes lies at the heart of clinical medicine as insights into how effective a particular treatment is able to cure a disease are based on this. Not surprisingly, most textbooks on clinical medicine and pathology contain extensive descriptions of how a disease progresses and likely reacts to particular treatments in the course of time. Yet, there has been very little research where these qualitative descriptions have been substantiated in a detailed, quantitative way. In research, the temporal dimension is usually only explored by describing the outcome of treatment after some time. One of the problems faced by researchers who wish to obtain such insight is the relatively small size of clinical datasets. Often, data concerns something from a hundred to a few hundreds of patients. However, the wish to develop a temporal model usually increases the demands for data, and as a consequence various simplifying assumptions have to be made.

A solution that is usually considered in clinical problems is to build a model that covers the entire time span, without distinguishing any of its time points [1–5]. Therefore, the model has the same properties for every instant, whether the second or the last one, e.g. as modeled by the well-known first-order homogeneous Markov chains [6]. A generalization of Markov chains to multivariate problems are dynamic Bayesian networks [7,8], DBNs for short, which are a family of models that has been applied to a number of real-world domains, such as medicine [9–12] and bioinformatics [13–15]. Such probabilistic graphical models allow to reason about the interactions of features of interest in an intuitive, temporal and compact fashion, while having a sound basis in probability theory. This will yield more robust models, making the use of these models attractive when dealing with small datasets. However, while it solves the robustness problem, it introduces an undesirable effect: there is no distribution specificity over the time series. Hence, one will never learn the details of the underlying process as was the aim in the first place.

* Corresponding author.
E-mail addresses: mbueno@cs.ru.nl (M.L.P. Bueno), arjenh@cs.ru.nl (A. Hommersom), peterl@cs.ru.nl (P.J.F. Lucas), m.lappenschaar@science.ru.nl (M. Lappenschaar), Joost.Janzing@radboudumc.nl (J.G.E. Janzing).

http://dx.doi.org/10.1016/j.jbi.2016.05.003
1532-0464/© 2016 Elsevier Inc. All rights reserved.

It is known that in many clinical situations the dependences between symptoms and signs change over time, as in the case of intervention studies, where different sets of correlations are naturally expected in the course of time, due to the nature of this kind of study. Hence, a temporal graphical model that is allowed to vary in structure and probability distribution as a function of time would capture these complex dynamics, providing a better fitting and more insightful model that really helps in understanding the underlying process. Although the notion of non-homogeneous time model is certainly not new, the innate technical difficulties to learn Bayesian network-based models for time series has forced most of the developed models to employ a number of approximations. Typically, these have been concentrated on biological processes, where regime shifts are assumed to be smooth [14,16,15]. It is difficult to suppose that these assumptions are naturally valid for processes with a different nature, where the variety of eligibility criteria and unexpected patient response to drugs can make the distribution regimes over time vary widely. Thus, a systematic algorithm that finds the appropriate time points to obtain new time-dependent parameters, taking into account the scarcity of data and the wish to obtain a robust model, is needed. To the best of our knowledge, this idea has never before been explored in learning Bayesian network-based models from data.

In this work, we propose a heuristic procedure to explore and learn over the space of non-homogeneous time dynamic Bayesian networks, taking into account the balance between specificity and simplicity. The approach starts with a fully homogeneous model, and incrementally replaces parts of it by sub-models that are valid for specific time periods. The increase of complexity is allowed if there is a two-part split of one of the current sub-models that is able to improve the fit over a training and test setting. The heuristic makes few assumptions regarding the process, the main one being the fact that the process duration is partitioned in the same way for every feature involved. We call this new kind of non-homogeneous time models *partitioned dynamic Bayesian networks*.

In order to demonstrate the applicability of the type of non-homogeneous time graphical models proposed here, an extensive set of simulations and real-world-based experiments were carried out. Simulating experiments showed that the heuristic algorithm was capable of constructing adequate solutions, in terms of goodness of fit and statistical distance to the original distributions, in consonance with the underlying processes that generated data, whether it was homogeneous or non-homogeneous. As a consequence, the advantages over homogeneous models as DBNs are highlighted when the underlying data generation process was not homogeneous. The experimental setup allowed to shed light on the behavior of the heuristic to learn proper models in the case of small datasets, which indicated that it tends to operate in a more conservative manner when dealing with these difficult situations, although still being capable of producing time-flexible and accurate models. Concerning more general settings, the experiments provided evidence that the greedy strategy has a proper behavior in the vast majority of simulations. Additionally, a study case on psychotic depression was conducted, where the models learned by the heuristic were discussed in detail. Then, the models learned for the psychiatry data were used to provide plausible answers for clinically relevant questions concerning this mental disorder, taking into account the time granularity of the original study for rendering predictions regarding symptom association on diverse future instants.

The remainder of this paper is organized as follows. Section 2 describes related literature of time homogeneous and non-homogeneous dynamic Bayesian networks in clinical and biological domains, followed by basic definitions concerning Bayesian networks in Section 3. The heuristic procedure to learn non-homogeneous time dynamic Bayesian networks is presented in Section 4.2. Simulations to evaluate the learning procedure are discussed in Section 5, while the models learned from psychiatry data (psychotic depression) are discussed in Section 6. Clinically-oriented discussions based on the psychiatry models are provided in Section 7, and lastly Section 8 gives the conclusions and suggestions for future research.

2. Related research

There has been quite some research on the application of Bayesian network models to the clinical domain. To a lesser extent, models that take time into account, such as dynamic Bayesian networks, have been considered in the past. Relevant research include obtaining problem insight by analyzing the structure and parameters of a DBN, and the use of DBN models for specific tasks such as diagnosis and prognosis. For example, the learned structure of a DBNs has been explored for finding correlations among different brain regions in several disorders, such as schizophrenia [3] and Alzheimer's disease [4]. These results have been used to confirm known correlations as well as to reveal new ones. Furthermore, the sensitivity of the influence of parameter variation in DBNs has been investigated in the context of ventilator-associated neumonia [17].

Another aspect of DBNs explored in the clinical domain is the predictive ability for several tasks, e.g. diagnosis [10,1] and prognosis [2]. An advantage of modeling stochastic processes using models as DBNs lies in the capability of producing updated predictions as new observations become available while the process progresses. This can be achieved by taking into account some form of patient history, producing potentially more accurate predictions. Real cases have shown the benefits of this type of multiple prediction, e.g. to diagnose ventilator-associated pneumonia [10]. The application of DBNs and similar models in clinical domains has been compared to similar formalisms in a recent survey [9].

Although DBNs have been reasonably studied for their capability to deal with clinical problems, this is not the case for more flexible models, e.g. when the time homogeneity assumption is rejected. These models address mainly the analysis of change in structure at individual time points, in the scope of a specific disease process [18]. On the other hand, more sophisticated models have been developed in other fields, mainly on biological processes [13,16,15,14]. In particular, these models are constructed certain assumptions often motivated by domain knowledge; for example, in some biological processes the intensity of interactions change over time, but no interaction is created or destroyed [16]. These models deviate from the partitioned dynamic Bayesian networks (PDBNs for short) that are proposed here, essentially in three points. Firstly, additional restrictions are usually imposed to the model structure, ranging from constrained intra-temporal interactions [14,15] to completely fixed structure with flexibility on the parameter space only [16]. Furthermore, neighboring homogeneous parts of the process are assumed to differ in a smooth fashion, while the family of PDBNs models does not impose this; the basic assumption in PDBNs is that the process duration is partitioned in the same way for every feature involved. Consequently, these two points imply that PDBNs indeed render a different class of models. The other distinction refers to the learning approach, which is based on sampling strategies in many works in bioinformatics [13–16], which in turn can depend on additional assumptions in order to be feasible. Earlier work on models to represent non-homogeneous stochastic processes include research on engineering problems [19].

Clearly, clinical problems are potentially prone to exhibit a temporal behavior that may be different from the biological processes studied so far. To illustrate this, consider the case of intervention studies, where specific criteria exist to define eligible patients. Consequently, imposing the previous assumptions on the manner in which pieces of the process evolve can forbid capturing the temporal dynamics accurately. Therefore, there is a need to define and construct models of non-homogeneous time in a systematic manner, which will be able to reveal more about the underlying structure of processes in clinical domains.

3. Preliminaries

Bayesian networks, BNs for short, provide a convenient and intuitive way to express probabilistic dependences and independences among random variables by means of a graphical structure. As such, they provide a compact representation of a joint distribution, which is crucial in real problems, e.g. in medicine, genetics, and climatology, where the number of variables and their domains can be considerably large. In particular, the manner in which parameters of a BN are elicited is local, in the sense that it involves the specification of a set of conditional distributions for each individual variable. This contrasts to a naive enumeration of the joint distribution which inevitably loses the opportunity to capture properties of the distribution. Such structural properties are advantageous when eliciting network parameters from experts.

In the following, we present some definitions and fix the notation. Each random variable (r.v.) is denoted by an upper-case letter, such as X, Y and Z; sets of random variables by bold face upper-case letters, e.g. $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$, while each value taken on by a r.v. by lower case letters, as in $val(X) = \{a, b, c\}$, where $val(X)$ denotes the domain of X. A set of random variables indexed by a time interval $[t_1, t_2]$ is denoted, for example, by $\mathbf{X}^{(t_1:t_2)}$. We employ the terms 'random variables' and 'nodes' of a graph interchangeably.

Furthermore, in a finite time series, a fixed set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, where each X_i is known as a feature, is observed repeatedly over a set of time points $t = 0, \ldots, T$. The set of features $\mathbf{X}$ is indexed by time t, resulting on a time series made of the variables $\mathbf{X}^{(0:T)}$. For each $i = 1, \ldots, n$, the variables $X_i^{(t)}$ consist of template variables of the feature X_i, i.e. they all have the same domain. We say that the duration or length of such time series is equal to $T + 1$ time points, where an instantiation of the variables $\mathbf{X}^{(0:T)}$ are referred to as a sequence of the time series. Hence, an instantiated time series correspond to a dataset containing a collection of sequences of that time series. In practice, in a dataset of sequences each random variable $X_i^{(t)}$ accommodates multiple values, where each one is an instantiation of the variable as observed on each sequence (e.g. a symptom measured at time t for multiple patients).

3.1. Bayesian networks and time

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a set of random variables, and $G = (\mathbf{X}, \mathbf{A})$ be a directed acyclic graph with node set $\mathbf{X}$ and arc set $\mathbf{A} \subseteq \mathbf{X} \times \mathbf{X}$. A *Bayesian network* $\mathcal{B}$ over $\mathbf{X}$ is a pair (G, P), such that the joint distribution of $\mathbf{X}$ factorizes over G, i.e.

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \pi(X_i)) \tag{1}$$

where $\pi(X_i)$ stands for the set of parents of node X_i in G, and P denotes a set of conditional probability distributions (CPDs), also known as conditional probability tables (CPTs) when the variables are finite. A consequence of the factorization of P in terms of G is that the independences that are codified in G must hold in P as well.

One of the most popular probabilistic models to handle time series are Markov chains (MCs) [6]. A *Markov chain* is a stochastic process consisting of a collection of random variables $(X^{(0)}, \ldots, X^{(T)})$, where all the variables have the same domain, usually known as the state space of the chain. The number of parameters needed to define the joint $P(X^{(0)}, \ldots X^{(T)})$ can be substantially large, what would forbid applying such model to many real applications, e.g. due to the large amounts of data required to learn such highly parameterized distribution. The typical solution for this issue is to assume two properties, namely, the Markovian property (also known as memoryless system property), and the time homogeneity property [8]. Consequently, the joint of a typical MC can be computed as:

$$P(X^{(0:T)}) = P(X^{(0)}) \prod_{t=0}^{T-1} P(X^{(t+1)} | X^{(t)}) \tag{2}$$

where $P(X^{(0)})$ is known as the *initial distribution*, and $P(X^{(t+1)} | X^{(t)})$ is known as the *transition distribution*. Note that a MC can be seen as a particular case of BNs, specifically a BN with structure $X^{(0)} \rightarrow \cdots \rightarrow X^{(T)}$.

In multivariate problems, a natural generalization of MCs are the dynamic Bayesian networks [20,8]. As such, DBNs model a multidimensional variable $\mathbf{X}$ over a time horizon $[0, T]$, where a variable at time $t + 1$ might be reached directly from variables at t and $t + 1$ only. Hence the transitions of a DBN can be seen as a local model of $\mathbf{X}^{(t)}$ conditioned on its parents, and therefore can be represented as a conditional Bayesian network. A *conditional Bayesian network* over $\mathbf{X}^{(t+1)}$ conditioned on $\mathbf{X}^{(t)}$ is a directed acyclic graph with node set $\mathbf{X}^{(t)} \cup \mathbf{X}^{(t+1)}$ that defines a conditional distribution

$$P(\mathbf{X}^{(t+1)} | \mathbf{X}^{(t)}) = \prod_{i=1}^{n} P(X_i^{(t+1)} | \pi(X_i)) \tag{3}$$

where $\pi(X_i) \subseteq \mathbf{X}^{(t)} \cup \mathbf{X}^{(t+1)}$.

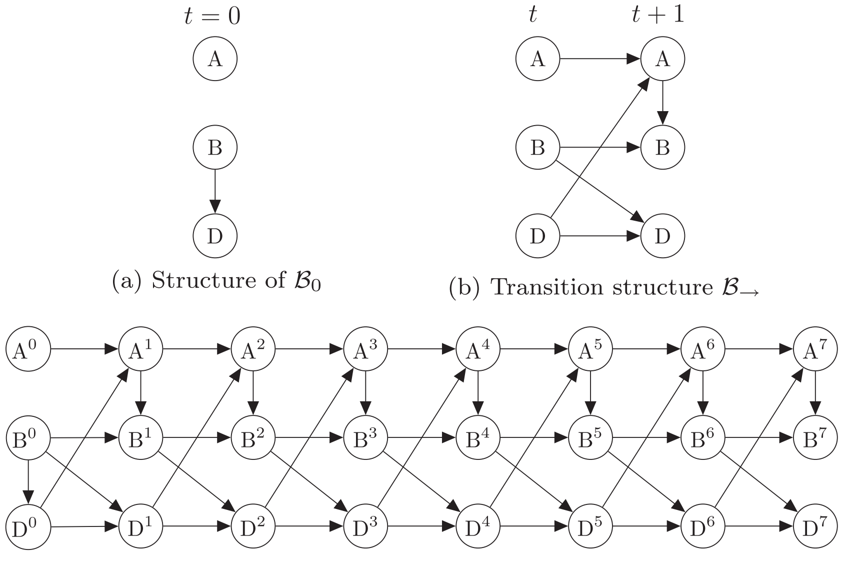
Since a DBN is Markovian, it follows that a single conditional BN is sufficient to capture the transition behavior over the entire process duration. Hence, a *dynamic Bayesian network* is defined as a dynamic system over $(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(T)})$, which initial distribution is given by a Bayesian network $\mathcal{B}_0$ over $\mathbf{X}^{(0)}$, and transition distribution is given by a conditional Bayesian network over $\mathbf{X}^{(t+1)}$ conditioned on $\mathbf{X}^{(t)}$. Putting these two pieces together produces an unrolled DBN, i.e. a Bayesian network with joint

$$P(\mathbf{X}^{(0:T)}) = \prod_{i=1}^{n} P(X_i^{(0)} | \pi(X_i, \mathcal{B}_0)) \prod_{t=0}^{T-1} \prod_{i=1}^{n} P(X_i^{(t+1)} | \pi(X_i, \mathcal{B}_{\rightarrow})) \tag{4}$$

where $\pi(X_i, \mathcal{B})$ denotes the parent set of X_i according to the structure of the BN $\mathcal{B}$. Note that the parent sets of each node are not indexed by time, since the distribution is time homogeneous.

The transition structure in a DBN associates to each node a parent set containing nodes from the previous instant (i.e. at t) and from the current instant (i.e. at $t + 1$). In the context of all nodes of a conditional BN, the set of arcs that go from t to $t + 1$ are called *inter-temporal arcs*, while the arcs from a variable at $t + 1$ to another variable at $t + 1$ are called *intra-temporal arcs*.

Example 3.1. In a disease process that lasts for 8 weeks, two symptoms (denoted by features A and B) and the administered drug quantity (denoted by feature D) are observed weekly. A DBN is used to model this problem, where the structures of the initial BN $\mathcal{B}_0$ and the conditional BN $\mathcal{B}\rightarrow$ that models transitions are shown at the top of Fig. 1. From these two structures, the corresponding unrolled DBN can be obtained, as shown at the bottom of Fig. 1.



(a) Structure of $\mathcal{B}_0$

(b) Transition structure $\mathcal{B}_\rightarrow$

(c) Unrolled DBN for the process duration.

Fig. 1. An example of DBN for a disease process that lasts over $[0, 7]$.

Example 3.2. Based on Example 3.1, probabilistic queries can be computed making use of the structure. If the initial symptoms and drug quantities are known for a patient at $t = 0$, one might be interested in a prediction for the symptom A two weeks further; therefore, the distribution $P(A^{(2)}|A^{(0)}, B^{(0)}, D^{(0)})$ must be computed. As $A^{(2)}$ is indirectly connected to the observed variables, the connecting variables are marginalized out according to the structure, resulting in

$$\begin{aligned} P(A^{(2)}|A^{(0)}, B^{(0)}, D^{(0)}) &= \Sigma_{A^{(1)}, D^{(1)}} P(A^{(2)}, A^{(1)}, D^{(1)}|A^{(0)}, B^{(0)}, D^{(0)}) \\ &= \Sigma_{A^{(1)}, D^{(1)}} P(A^{(1)}|A^{(0)}, D^{(0)}) P(D^{(1)}|B^{(0)}, D^{(0)}) \\ &\quad \times P(A^{(2)}|A^{(1)}, D^{(1)}) \end{aligned}$$

If new evidence becomes available, an updated distribution is computed, e.g. as in $P(A^{(2)}|A^{(1)}, B^{(1)}, A^{(0)}, B^{(0)}, D^{(0)})$. Interestingly, $B^{(1)}$ is irrelevant for the prediction of $A^{(2)}$ on both queries, as deduced from the structure.

3.2. Psychotic depression

In Section 6, the value of the techniques developed in this paper is demonstrated by data from patients whom have been treated for psychotic depression. Psychotic depression is a mental disorder in which depressive and psychotic symptoms are present. Due to the psychotic symptoms, as hallucinations and delusions, psychotic depression is considered a severe subtype of Major Depressive Disorder (MDD). The interaction between these two kinds of symptoms makes the clinical treatment challenging, often requiring a combination of antidepressant and anti-psychotic drugs [21]. Due to its complex, dynamic nature, the temporal analysis of the interaction between the diverse symptoms and signs supports the insight for determining an adequate treatment of the patient. Because of the complexity, it is also valuable to determine the potential future trajectories of a patient, after observing the current symptoms. These kind of practical questions can be properly formulated and answered with different types of temporal Bayesian networks, as DBNs and partitioned DBNs introduced in the next section.

Quantitatively, depression can be measured by means of rating scales, such as the *Hamilton depression rating scale* (HDRS17), which is a 17-item scale that accounts for several aspects of depression [22]. Each item of the scale assumes an integer value, starting from 0 and increasing proportionally to the intensity of the symptom. The final measurement of depression can, then, consist of the total sum of the 17 items, or the sum of a selection of part of the HDRS, e.g. the melancholia sub-scale [23]. In some cases, symptoms can be considered individually as well. On the other hand, features of psychosis are considered dichotomized variables, taking values on 0 if the symptom is absent, and 1 otherwise. Whereas the Hamilton depression rating scale appears to be useful as a tool for the psychiatrist, it lacks the ability to support the development of insight into the dynamics of the disease.

4. Partitioned dynamic Bayesian networks

Models of non-homogeneous time are defined from a set of transition distributions, followed by associating them to a partition of the duration of a time series. The terminology that is used to describe this type of model in the literature is not uniform [14,16,15], added to the fact that the definition itself of what is a non-homogeneous time model is inconstant as well. In this work, the central idea relies on making the dependence on time by partitioning the time series duration and associating each part to a homogeneous model, i.e. a DBN valid within a sub-range of the time series. As such, we refer to this class of models as *partitioned dynamic Bayesian networks*. We proceed in the following towards a formalization of PDBNs, its associated concepts, and lastly a procedure to learn PDBNs by exploring the search space heuristically.

4.1. Model specification

Let a *time partition* of a set of integers $\{0, 1, \ldots, T\}$ be a set of $k > 0$ intervals $(0, t_1], (t_1, t_2], \ldots, (t_{k-1}, t_k]$, such that (1) $t_1 > 0$ and $t_k = T$; and (2) $i > j$ implies that $t_i > t_j$, for every $i, j = 1, \ldots, k$. If $(t_i, t_j]$ is one of these intervals, denote by t_j the *cut* of the interval. Hence, a time partition has a cut set of the form $\{t_1, t_2, \ldots, t_k\}$. These notions are useful to combine the partitioning of a temporal process with distribution transitions, as follows.

Let $(\mathcal{B}_1,\ldots,\mathcal{B}_k)$ be a collection of conditional BNs over $\mathbf{X}^{(t+1)}$ conditioned on $\mathbf{X}^{(t)}$. As such, each i-th conditional BN has a set of CPDs P_i. Consider a time partition of $[0,T]$ into k cuts; let each cut $i=1,\ldots,k$ be associated to a $\mathcal{B}_i$. A *partitioned dynamic Bayesian network with k cuts*, denoted by PDBN-k, is a dynamic system over $(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(T)})$ in which the transition distribution for every pair of time points $(t,t+1)$ is given by the $\mathcal{B}_i$ associated with the cut that contains such pair. We use the term *distribution cut* to denote a cut in the context of a PDBN. It follows from this definition that a DBN is a PDBN with a single cut at $\{T\}$, and hence, it is a PDBN-1.

The joint distribution of an unrolled PDBN with k cuts $\{t_1,t_2,\ldots,t_k\}$ can be obtained from the previous definitions and assumptions as:

$$P(\mathbf{X}^{(0:T)}) = \prod_{i=1}^{n} P_0(X_i^{(0)}|\pi(X_i,\mathcal{B}_0)) \prod_{r=1}^{k}\prod_{t=t_{r-1}}^{t_r-1}\prod_{i=1}^{n} P_r(X_i^{(t+1)}|\pi(X_i,\mathcal{B}_r)) \qquad (5)$$

where $t_0=0$ and P_r refers to the CPDs pertaining to the conditional BN $\mathcal{B}_r$. Note that the parent set of each X_i is coupled to a $\mathcal{B}_r$ in $\pi(X_i,\mathcal{B}_r)$, and thus no time indexing is needed, since within the scope of each conditional BN the process is homogeneous.

Example 4.1. Consider the disease process of Example 3.1. A PDBN-2 is defined for this problem, consisting of two cuts $\{2,7\}$, which initial structure $(\mathcal{B}_0)$ and transition structures are shown on Fig. 2. Each cut of the PDBN is associated to a conditional BN, namely, $\mathcal{B}_{0\to 2}$ dictates the transitions in $\{0,\ldots,2\}$, and $\mathcal{B}_{2\to 7}$ the transitions in $\{2,\ldots,7\}$.

Unrolling this PDBN-2 for the process duration yields the joint

$$\begin{aligned} P(\mathbf{X}^{(0:7)}) &= \prod_{i} P_0(X_i^{(0)}|\pi(X_i,\mathcal{B}_0)) \prod_{0\leqslant t\leqslant 1}\prod_{i} P_{0\to 2}(X_i^{(t+1)}|\pi(X_i,\mathcal{B}_{0\to 2})) \\ &\quad\times \prod_{2\leqslant t\leqslant 6}\prod_{i} P_{2\to 7}(X_i^{(t+1)}|\pi(X_i,\mathcal{B}_{2\to 7})) \end{aligned} \qquad (6)$$

where $\mathbf{X}=\{A,B,D\}, X_i\in\mathbf{X}$, and $P_{i\to j}$ refers to the CPDs pertaining to the conditional BN $\mathcal{B}_{i\to j}$.

4.2. A heuristic search procedure

In this section, we present a heuristic algorithm to build PDBNs in an incremental fashion from a dataset of sequences. As in many clinical studies there is typically a scarcity of data, mainly in terms of number of sequences (e.g. represented by patients), the central idea of the procedure is to prefer less complex models. In order to achieve this, the heuristic assumes that a proper criterion for model selection that prevents overfitting is used, which is naturally dependent on the application domain and characteristics of the data. Hence, when constructing a model, the heuristic iteratively increases the complexity as long as it is beneficial for its score; if adding complexity is not, the procedure stops adding further complexity. Additionally, the procedure has a hill-climbing behavior by not further exploiting previous less complex solutions that were less promising when analyzed by the algorithm.

4.2.1. Algorithm description

Taking the aforementioned factors into account, we designed a procedure that starting from a DBN follows a sequence of incremental refinements to evolve it into a more specialized model. A refinement corresponds to split one of the transition distributions of the current PDBN. At each iteration a new cutting point is added without destroying the cuts already found previously. Consequently, the procedure is greedy since it does not explore the branching of solutions that are less interesting at each iteration. Nevertheless, it is important to consider that the strategy to search over the space of PDBNs is crucial, since the number of possible manners in which a discrete time series can be partitioned is computationally intractable. In order to be flexible, the complexity of the produced models can be controlled, as it is an input parameter of the algorithm.

The heuristic algorithm to learn PDBNs is presented in Algorithm 1. In order to be generic for different scoring criteria used to construct and evaluate PDBNs, we emphasize the search for cut sets instead of PDBNs explicitly. The algorithm starts with the current best cut set as the singleton $C=\{T\}$, which stands for a fully homogeneous model. Let us denote by s the size of the current cut set. Entering the outer loop (Line 2) will first evaluate new cut sets with size $s+1$, each one consisting of the current C unified with a new cut that does not exist in C (Line 3). After finishing the inner loop, it is verified whether the current iteration has found an improved cut set, i.e. a cut set whose evaluation is better than C. In case positive, C is replaced by the best cut set among those (Lines 5 and 6). The algorithm continues this incremental construction of cut sets while the current iteration is capable of producing a new cut set with size $(s+1)$ that overcomes the current C and the maximum number of cuts (the input parameter k) is not reached. At the end (Line 8), the heuristic returns the PDBN-k' learned from the best cut set found, where $k'\leqslant k$.

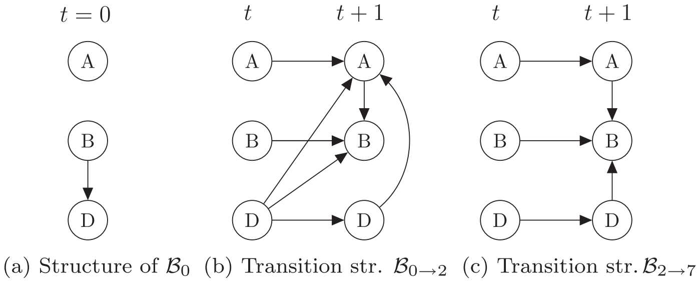
Algorithm 1. Builds a PDBN

```
  Input: $D$, a dataset of sequences with length $(0,\ldots,T)$;
  $k$, the maximum size of the cut set, $1\leqslant k\leqslant T$.
  Output: a PDBN-$k'$, where $k'\leqslant k$.
1: $C\leftarrow\{T\}$, where $s=|C|$.
2: while $|C|<k$ do
3:     For each $c\in\{1,\ldots,T\}-C$, construct a new cut set
       $C\cup\{c\}$. Denote the new cut sets by $\mathbf{C}=\{C_1,\ldots,C_r\}$, where
       $1\leqslant r\leqslant T-|C|+1$ and $1\leqslant i\leqslant r$.
4:     Evaluate each cut set in $\mathbf{C}$ by means of a criterion $f$.
5:     if there is a new cut set $C_i\in\mathbf{C}$ such that $f(C_i)>f(C)$
       then
6:         Assign to $C$ the $C_j$ that maximizes $\{f(C_1),\ldots,f(C_r)\}$.
7:     else break the loop.
8: Return the PDBN-$k'$ learned from the cut set $C$, where $k'=s$.
```

4.2.2. Evaluation criterion

As Algorithm 1 shows, the criterion f abstracts the learning of PDBNs. This is motivated by the fact that choosing a proper evaluation strategy depends on the application and the characteristics of the data, which makes it difficult to set a single criterion that works best for all problems [24]. Generally speaking, a multitude of model selection criteria can be employed to determine how f is concretely implemented; some well-known criteria include cross-validation (e.g. based on the model's likelihood) and information theory criteria (e.g. Akaike information criterion – AIC – and Bayesian information criterion – BIC) [25]. For example, employing AIC would correspond to implement Line 3 first learning a PDBN using the full dataset (i.e. all the sequences), then computing the AIC of the resulting model over the full dataset as well, which value would correspond to f. Similarly, each sub-DBN can be learned using a standard Bayesian-network learning algorithm, typically falling within search-and-score and constraint-based approaches [8].

An important criterion for the present purpose of learning dynamic models from a small dataset is to avoid overflexible models that may overfit and reflect high variance in model learning. Five- or ten-fold cross-validation overestimates prediction error



(a) Structure of $\mathcal{B}_0$ (b) Transition str. $\mathcal{B}_{0\to 2}$ (c) Transition str. $\mathcal{B}_{2\to 7}$

Fig. 2. An example of PDBN-2. Each transition structure is denoted by $\mathcal{B}_{i\to j}$, i.e. it governs each transition from t to $t+1$ in $\{i,\ldots,j\}$. Nodes on the left and the right side occur at t and $t+1$ respectively, except for the initial BN.

and thus punishes overfitting. As a consequence, it can be combined with likelihood for model selection and a method to punish model complexity, as for example provided in AIC and BIC, is not required [26].

4.2.3. Complexity

Initially, the cut set maintained by the algorithm is $C=\{T\}$. At the first iteration of the outer loop, new cut sets with size $s+1$ are built, consisting of C plus a new element; there are $T-1$ manners to make this inclusion. At the second iteration, there are $T-2$ possible cut sets to be constructed, and so on, until the last iteration, in which there is only one cut to be inserted in the current C. Thus, the total number of cut sets constructed by the heuristic is in $\mathcal{O}(T^2)$, considering the worst case.

Naturally, the dominant part of the heuristic's total cost corresponds to learning models. In the case of learning DBNs, the input can be seen as a transition dataset $(\mathbf{X},\mathbf{X}')$, consisting of all the data $(\mathbf{X}^{(i)},\mathbf{X}^{(i+1)})$, $i=0,\ldots,T-1$, merged, as if there were a single time transition. Note that this construction is sound since the model is time-homogeneous. If the original dataset D consists of m sequences (of length $T+1$), this merged dataset will consist of mT short sequences (of length 2). Thus, abstracting the cost of learning a DBN by means of a cost function g will lead to a cost of $\mathcal{O}(g(mT))$ to learn a DBN. The case of learning PDBNs-$k, k>1$, can be seen as learning k sub-DBNs made of potentially different number of sequences, as dictated by the cut set of the PDBN. Note that when the number of cuts is maximal, it implies learning T sub-DBNs, each one from a transition dataset $(\mathbf{X}^{(i)},\mathbf{X}^{(i+1)})$ consisting of m short sequences. As each of these sub-DBNs would cost $g(m)$, learning such PDBN would require $\mathcal{O}(Tg(m))$. Thus, the cost dynamics of the heuristic moves between these two cost estimations as the size of the cut set changes.

5. Empirical evaluation via simulations

5.1. Simulation parameters

We considered experiments with simulated data to obtain a more general assessment of the proposed method for learning PDBNs[1]. Time series with varying length and number of sequences were generated, resulting in diversified datasets. Specifically, we considered the number of features, denoted by $n\in\{2,6,10,14,18\}$, where each time series is composed by sequences with length of 10 or 30 points. Hence, simulations accounted for time series ranging from 20 to 540 random variables in total. For each n and time series length, datasets were randomly generated containing different number of sequences, denoted by $d\in\{100,500,2000,5000\}$. Thus, simulations allowed a reasonable evaluation in terms of different feature spaces and dataset sizes.

On each simulation, a random DBN or PDBN-k was constructed, consisting of n binary features per instant t. Structurally, a random PDBN-k consists of k random sub-DBNs, where each random sub-DBN had its graphical structure uniformly generated at random [27], and distribution parameters determined randomly as well (no noise was introduced in the model's parameters). Hence, each node of an unrolled PDBN assumed a Bernoulli distribution. Given a random PDBN-k and a random cut set of length k, whose last cut corresponds to the length of the sequences that are to be sampled from the model, four disjunct datasets were constructed, one for each value of d. In other words, a common underlying model was used for each group of simulations since the experiments also aimed at studying the effect on the heuristic's capabilities over different quantities of data.

Each dataset was generated from either a DBN or a PDBN (i.e. both random structure and parameters). The initial aim is to verify experimentally whether the construction algorithm is able to learn the adequate "kind" of model, wrt. the reference model (a random DBN or PDBN) used to simulate data. Moreover, the cuts of the learned models were compared to the cuts of the reference models, where we use the following notation: if the cuts of the reference and learned models are equal, we write $=$; if the cuts of the learned model include all the cuts of the reference one, we write $\subseteq +a$, where a denotes the number of additional cuts included by the learned model; finally, if none of these criteria is met, we write $\not\subseteq$. Although these two criteria are useful to perform a structural comparison in terms of the number and position of distribution cuts, they obviously do not provide information about the distance between the probability distributions of two models. To this end, the Kullback–Leibler divergence [25] between the marginal distribution of each feature at each t, denoted by $X_i^{(t)}$, was considered. Specifically, on each simulation, it is computed $\sum_{i,t}\mathrm{KL}(P(X_i^{(t)})\|Q(X_i^{(t)}))$, for all $i=1,\ldots,n$, and $t=0,\ldots,T$, i.e. the sum of the divergences between the marginal distributions as dictated by two distributions P and Q, in this case a reference distribution and a learned distribution respectively. Note that this criterion differs from the usual KL divergence over the entire joint distribution, which is computationally prohibitive for most of the simulations covered in this section. As usual, the KL divergence should be minimized, reflecting the amount of extra information needed to codify P using Q.

5.2. Learning and evaluating PDBNs

In the experiments of this paper, we implemented the evaluation criterion by means of a 10-fold cross-validation. Cross-validation minimizes the effect of overfitting (see Section 4.2.2); we describe the procedure in detail in the following. Let

[1] The implementation is available at http://www.cs.ru.nl/~mbueno.

$C_i = \{t_1, \ldots, t_k\}$ be a cut set of a time series over $[0, T]$; in the context of Algorithm 1, C_i corresponds to the a new cut set that is built in Line 3. For each cross-validation fold, the training data is used to learn a PDBN-k with cut set C_i, while the test data is used to compute the log-likelihood of such PDBN. After processing all the folds, the mean of the log-likelihoods is taken, which represents the evaluation value of the cut set C_i, indicated in Algorithm 1 by $f(C_i)$. Finally, when deciding between two cut sets (e.g. as in Line 5), the algorithm chooses the one having the higher mean log-likelihood. Once the heuristic search has finished (i.e. after leaving the outer loop of Algorithm 1), and the best cut set has been determined, a PDBN with such cut set is learned using the full dataset, i.e. both training and test data. Such PDBN corresponds to the output of the procedure.

In order to learn a PDBN-k, k homogeneous models are learned using the corresponding portions of the training data according to its cut set. Each of the k sub-DBNs is learned separately. As it happens with Bayesian networks learning, typically search-and-score and constraint-based methods are used during learning; in the experiments of this section and Section 6 the AIC criterion was employed to score each sub-DBN, which means yielding a score proportional to the likelihood of the model and a penalization term for the complexity. This is the case for example when showing the resulting model in the psychiatry case (Section 6).

5.3. Results and discussion

The results of simulations with data generated from DBNs, PDBNs-2 and PDBNs-3 models are shown on Tables 1–3 respectively. Note that a DBN was learned on every case to serve as a baseline method, specially when simulating data from non-DBNs; the ability of the learned DBNs are indicated on the sixth column of the tables. The models learned by the heuristic from DBN data (Table 1) had structural partitioning in accordance with the reference models on most cases, showing that the heuristic was capable of retrieving the adequate type of model. When the returned models were not a DBN, these were mostly slightly more complex ones (i.e. PDBNs-2). Interestingly, the KL divergence between the learned PDBNs and the respective reference models are comparable to the divergence of the learned DBNs, i.e. although consisting of additional transition distributions, the learned PDBNs captured the reference distribution as good as the learned DBNs did.

The models returned by the heuristic based on data produced by PDBNs-2 and PDBNs-3 (Tables 2 and 3) support analogous points discussed just before. Furthermore, these tables show that the KL divergences of the PDBNs learned heuristically were substantially lower than those of the learned DBNs, i.e. the former are closer to the reference ones. This fact was more prominent when the length of the time series was increased to 30. Intuitively, a DBN "averages" over the distribution underlying data; if most of the transitions were originated from a single distribution, then the few remaining ones will tend to have less impact on the distribution learned by the DBN. On the PDBNs-2 and PDBNs-3 cases where the first cut was situated around half of the sequence duration, there were at least two different transition patterns, what tends to make a DBN less representative of each individual transition. Overall, it is worth noting that the cases where the heuristic procedure was not capable of constructing models with the same structural partition of transitions as the reference models do have some particularities. Namely, these cases contain little features (mostly $n = 2$) or have few sequences. Despite not returning the exact type of model, the KL divergences of these PDBNs were noticeably smaller than the divergences of the learned DBNs, suggesting that the heuristic is capable of finding alternative "routes" with good fit in unfavorable scenarios.

Table 1
Simulations with data generated by DBNs, where n and d denote the number of features and the number of sequences respectively. The heuristically learned models, the reference models and the learned DBNs are abbreviated as *Heur*, *Ref*, and *Le. DBN* respectively.

n	d	Model (Heur.)	Cut sets (Ref. and Heur.)	Cuts diff.	$\sum$ KL (Le.DBN)	$\sum$ KL (Heur.)
Time series length = 10						
2	100	DBN	(9)	$=$	0.04	0.04
2	500	DBN	(9)	$=$	0.01	0.01
2	2000	DBN	(9)	$=$	0	0
2	5000	PDBN-2	(9); (7,9)	$\subseteq +1$	0	0
6	100	DBN	(9)	$=$	0.17	0.17
6	500	DBN	(9)	$=$	0.04	0.04
6	2000	DBN	(9)	$=$	0.01	0.01
6	5000	DBN	(9)	$=$	0.01	0.01
10	100	DBN	(9)	$=$	0.24	0.24
10	500	DBN	(9)	$=$	0.09	0.09
10	2000	DBN	(9)	$=$	0.02	0.02
10	5000	DBN	(9)	$=$	0.02	0.02
14	100	DBN	(9)	$=$	0.38	0.38
14	500	DBN	(9)	$=$	0.07	0.07
14	2000	DBN	(9)	$=$	0.03	0.03
14	5000	DBN	(9)	$=$	0.02	0.02
18	100	DBN	(9)	$=$	0.23	0.23
18	500	DBN	(9)	$=$	0.07	0.07
18	2000	DBN	(9)	$=$	0.03	0.03
18	5000	DBN	(9)	$=$	0.02	0.02
Time series length = 30						
2	100	DBN	(29)	$=$	0.01	0.01
2	500	DBN	(29)	$=$	0.01	0.01
2	2000	DBN	(29)	$=$	0	0
2	5000	PDBN-2	(29); (1,29)	$\subseteq +1$	0.01	0.01
6	100	DBN	(29)	$=$	0.16	0.16
6	500	DBN	(29)	$=$	0.03	0.03
6	2000	DBN	(29)	$=$	0.02	0.02
6	5000	DBN	(29)	$=$	0.02	0.02
10	100	DBN	(29)	$=$	0.13	0.13
10	500	DBN	(29)	$=$	0.04	0.04
10	2000	DBN	(29)	$=$	0.03	0.03
10	5000	DBN	(29)	$=$	0.02	0.02
14	100	DBN	(29)	$=$	0.26	0.26
14	500	DBN	(29)	$=$	0.07	0.07
14	2000	DBN	(29)	$=$	0.04	0.04
14	5000	DBN	(29)	$=$	0.04	0.04
18	100	DBN	(29)	$=$	0.3	0.3
18	500	DBN	(29)	$=$	0.08	0.08
18	2000	DBN	(29)	$=$	0.05	0.05
18	5000	DBN	(29)	$=$	0.04	0.04

A summary of the results presented in Tables 1–3 is given in Table 4. Each row of the table aggregates simulations of DBNs, PDBNs-2 and PDBNs-3 according to the number of features and sequence length.

5.4. Small datasets

Finally, the simulations showed that the models learned by the heuristic from the smallest datasets (i.e. those containing 100 sequences) were simpler than the reference models used to generate simulated data on virtually every case. Hence, it was evidenced experimentally that the heuristic tends to operate in a conservative mode when there is scarcity of data; in other words, no overfitting was observed on these simulations.

With regard to the structural partitioning and quality measurements for these models: (1) the cuts of the learned models were all part of the cut sets of the reference models, on almost all cases (note that this includes all the cases with a $\not\subseteq$); and (2) the divergences of the learned PDBNs were substantially smaller than those of DBNs, specially when data was generated from PDBNs-3, indicating a decent learning ability of the heuristic in the difficult situation of small datasets.

Table 2
Simulations with data generated by PDBNs-2. For each case, the best values on the KL divergences are given in bold and followed by an asterisk.

n	d	Model (Heur.)	Cut Sets (Ref. and Heur.)	Cuts diff.	$\sum$ KL (Le.DBN)	$\sum$ KL (Heur.)
Time series length = 10						
2	100	PDBN-4	(1,9); (1,2,4,9)	$\subseteq +2$	0.18	**0.08***
2	500	PDBN-2	(1,9)	=	0.16	**0.02***
2	2000	PDBN-4	(1,9); (1,5,7,9)	$\subseteq +2$	0.19	**0.01***
2	5000	PDBN-4	(1,9); (1,5,8,9)	$\subseteq +2$	0.19	**0.01***
6	100	PDBN-2	(6,9)	=	1.88	**0.14***
6	500	PDBN-2	(6,9)	=	1.81	**0.03***
6	2000	PDBN-2	(6,9)	=	1.76	**0.01***
6	5000	PDBN-2	(6,9)	=	1.76	**0.01***
10	100	DBN	(8,9); (9)	$\not\subseteq$	1.06	1.06
10	500	PDBN-2	(8,9)	=	0.96	**0.05***
10	2000	PDBN-2	(8,9)	=	0.95	**0.02***
10	5000	PDBN-2	(8,9)	=	0.95	**0.01***
14	100	PDBN-2	(3,9)	=	3.07	**0.37***
14	500	PDBN-2	(3,9)	=	2.68	**0.1***
14	2000	PDBN-2	(3,9)	=	2.39	**0.03***
14	5000	PDBN-2	(3,9)	=	2.35	**0.02***
18	100	DBN	(1,9); (9)	$\not\subseteq$	1.57	1.57
18	500	PDBN-2	(1,9)	=	1.04	**0.09***
18	2000	PDBN-2	(1,9)	=	0.93	**0.02***
18	5000	PDBN-2	(1,9)	=	0.78	**0.02***
Time series length = 30						
2	100	PDBN-2	(15,29)	=	5.05	**0.09***
2	500	PDBN-2	(15,29)	=	5.05	**0.02***
2	2000	PDBN-7	(15,29); (2,6,15,20,26,28,29)	$\subseteq +5$	5.07	**0.02***
2	5000	PDBN-4	(15,29); (10,15,25,29)	$\subseteq +2$	5.08	**0.01***
6	100	PDBN-2	(18,29)	=	15.8	**0.12***
6	500	PDBN-2	(18,29)	=	15.7	**0.04***
6	2000	PDBN-2	(18,29)	=	15.76	**0.02***
6	5000	PDBN-2	(18,29)	=	15.95	**0.02***
10	100	PDBN-2	(20,29)	=	7.24	**0.26***
10	500	PDBN-2	(20,29)	=	7.25	**0.12***
10	2000	PDBN-2	(20,29)	=	7.2	**0.06***
10	5000	PDBN-2	(20,29)	=	7.13	**0.03***
14	100	PDBN-2	(21,29)	=	9.29	**0.35***
14	500	PDBN-2	(21,29)	=	9.09	**0.09***
14	2000	PDBN-2	(21,29)	=	9.09	**0.06***
14	5000	PDBN-2	(21,29)	=	9.02	**0.04***
18	100	PDBN-2	(17,29)	=	13.02	**0.34***
18	500	PDBN-2	(17,29)	=	12.82	**0.1***
18	2000	PDBN-2	(17,29)	=	12.57	**0.06***
18	5000	PDBN-2	(17,29)	=	12.64	**0.05***

Table 3
Simulations with data generated by PDBNs-3.

n	d	Model (Heur.)	Cut sets (Ref. and Heur.)	Cuts diff.	$\sum$ KL (Le.DBN)	$\sum$ KL (Heur.)
Time series length = 10						
2	100	PDBN-2	(1,6,9); (6,9)	$\not\subseteq$	2.73	**0.26***
2	500	PDBN-4	(1,6,9); (1,2,6,9)	$\subseteq$ +1	2.8	**0.02***
2	2000	PDBN-3	(1,6,9)	=	2.79	**0.01***
2	5000	PDBN-4	(1,6,9); (1,4,6,9)	$\subseteq$ +1	2.78	**0***
6	100	PDBN-3	(2,6,9)	=	3.37	**0.25***
6	500	PDBN-3	(2,6,9)	=	3.09	**0.04***
6	2000	PDBN-3	(2,6,9)	=	2.91	**0.02***
6	5000	PDBN-3	(2,6,9)	=	2.94	**0.01***
10	100	PDBN-2	(6,8,9); (6,9)	$\not\subseteq$	2.99	**1.83***
10	500	PDBN-3	(6,8,9)	=	2.85	**0.07***
10	2000	PDBN-3	(6,8,9)	=	2.8	**0.02***
10	5000	PDBN-3	(6,8,9)	=	2.78	**0.02***
14	100	PDBN-2	(2,3,9); (3,9)	$\not\subseteq$	6.5	**1.61***
14	500	PDBN-3	(2,3,9)	=	5.41	**0.1***
14	2000	PDBN-3	(2,3,9)	=	4.96	**0.04***
14	5000	PDBN-3	(2,3,9)	=	4.76	**0.02***
18	100	DBN	(1,8,9); (9)	$\not\subseteq$	2.17	2.17
18	500	PDBN-3	(1,8,9)	=	1.85	**0.1***
18	2000	PDBN-3	(1,8,9)	=	1.7	**0.03***
18	5000	PDBN-3	(1,8,9)	=	1.52	**0.02***
Time series length = 30						
2	100	PDBN-3	(15,17,29)	=	1.97	**0.1***
2	500	PDBN-3	(15,17,29)	=	1.93	**0.02***
2	2000	PDBN-6	(15,17,29); (1,6,15,17,22,29)	$\subseteq$ +3	1.92	**0.02***
2	5000	PDBN-5	(15,17,29); (3,15,16,17,29)	$\subseteq$ +2	1.92	**0.01***
6	100	PDBN-3	(18,19,29); (17,19,29)	$\not\subseteq$	17.15	**1.53***
6	500	PDBN-3	(18,19,29)	=	17.09	**0.05***
6	2000	PDBN-3	(18,19,29)	=	17.13	**0.03***
6	5000	PDBN-3	(18,19,29)	=	17.12	**0.02***
10	100	PDBN-3	(20,24,29)	=	25.57	**0.38***
10	500	PDBN-3	(20,24,29)	=	25.69	**0.07***
10	2000	PDBN-3	(20,24,29)	=	25.59	**0.05***
10	5000	PDBN-3	(20,24,29)	=	25.09	**0.03***
14	100	PDBN-3	(8,21,29)	=	15.53	**0.47***
14	500	PDBN-3	(8,21,29)	=	15.3	**0.17***
14	2000	PDBN-3	(8,21,29)	=	15.15	**0.07***
14	5000	PDBN-3	(8,21,29)	=	15.05	**0.04***
18	100	PDBN-3	(1,17,29)	=	12.63	**0.61***
18	500	PDBN-3	(1,17,29)	=	12.07	**0.11***
18	2000	PDBN-3	(1,17,29)	=	12.03	**0.06***
18	5000	PDBN-3	(1,17,29)	=	11.97	**0.05***

Table 4
Summary of simulations with DBNs and PDBNs. The second column refers to the average of $\sum$ KL(DBN) − $\sum$ KL(PDBN), hence positive values indicate higher divergences of DBNs. The third column shows the average of $\sum$ KL(PDBN). The fifth and sixth columns show the mean number of additional cuts (when a $\subseteq +a$ occurs.) and the number of $\not\subseteq$ occurrences.

n	DBNs-PDBNs	PDBNs	Equal cut sets (total)	Additional cut sets	Other cut sets (total)
Time series length = 10					
2	0.95	0.04	5(12)	1.5	1(12)
6	1.58	0.06	12(12)	0	0(12)
10	1.02	0.29	10(12)	0	2(12)
14	2.49	0.23	11(12)	0	1(12)
18	0.63	0.36	10(12)	0	2(12)
Time series length = 30					
2	2.31	0.03	7(12)	2.6	0(12)
6	10.82	0.17	11(12)	0	1(12)
10	10.81	0.1	12(12)	0	0(12)
14	8.02	0.14	12(12)	0	0(12)
18	8.2	0.15	12(12)	0	0(12)

6. Learning temporal models of psychotic depression

6.1. Bayesian networks in psychiatry

The use of probabilistic graphical models in psychiatry has been fairly narrow. Existing research is mainly restricted to semi-automatic and fully handcrafted approaches, namely, learning only the parameters from data [28,5] and eliciting both structure and parameters from descriptive statistics and expert knowledge [29,30]. Although making use of expert knowledge might be necessary, e.g. in order to include established medical knowledge, the use of a data-driven approach has been able to discover new and unexpected insights in a multitude of fields. Furthermore, an advantage of BN models that can be of interest in psychiatry studies lies on making predictions when provided with incomplete evidence (e.g. only a few symptoms). This feature has been explored in some studies [29,30], however at the individual level of a few patients (whether real or artificial), consequently, there is still a need for understanding associations between different variables in a more comprehensive and systematic way. This can include inferences for a population of patients, in order to reveal more general knowledge about, for example, the predictive power among different sets of features.

In particular, among the literature on BNs in psychiatry, time has not been a factor taken into account in a systematic manner so far. Except for the case that considers only the begin and the end of a treatment [28], research that takes into account the real time granularity has not been developed to this moment. A few important instances can be controlled treatments and longitudinal diagnosis, where the examination of some form of history or time series measurements would allow a more global comprehension of, for example, the evolution of mental illnesses and a more accurate and detailed diagnosis. Even in the absence of some of the features as input, models that capture a time dimension, as dynamic Bayesian networks [8], are still able to deliver predictions about future instants, what can be difficult for methods rooted on regression, for example. Furthermore, these can be regarded from a single point (e.g. the time point after the last measurement of the features) to multiple future time points. Besides prediction, temporal models can also be used to find associations taking into account the time dimension.

Within the field of psychiatry, diseases that have been covered under a BN approach include depression [28,29,31], social anxiety [5], schizophrenia [30], as well as analyzing the use of BNs on diagnosis in psychiatry [32]. Moreover, there is a lack of research of temporal models on psychotic depression, which besides being a severe mental disorder, brings an additional complexity due to the presence of both psychotic and depressive symptom factors.

6.2. Problem description and data preprocessing

To illustrate the use of non-homogeneous time probabilistic models and the heuristic construction procedure proposed in this work, we selected a case in psychiatry. It comprises a dataset from an original study designed to compare the efficacy of three drug treatment strategies (imipramine, venlafaxine and venlafaxine-quetiapine) in a sample of patients with psychotic depression over 7 weeks [33]. The primary outcome of the study aimed at investigating which strategy allowed superior reductions of depressive and psychotic symptoms at treatment endpoint. In this work, in turn, we aim at answering a different research question: *to which extent do depressive and psychotic symptoms interact over time*? To this end, temporal models as DBNs and PDBNs are shown to be adequate since a large range of hypothesis can be verified supported by those, while modeling explicit relationships between psychotic and depressive symptoms. We first discuss the results obtained by the heuristic algorithm when applied over psychiatry data, aiming at providing: (1) a more technical perspective based on fitting assessment between DBNs and PDBNs and (2) an investigation of the dependences in the graphical structure. Then, in Section 7 we make use of the obtained models to clinically-oriented research questions, as the one mentioned earlier.

Differently from the original study, in which the primary outcome was the sum of the 17-item Hamilton depression rating scale (HDRS17, see Section 3.2) [22], in this work we considered the individual symptoms of the HDRS17. The dataset consists of 122 patients' data, from which 100 are patients that completed the treatment. Given the limited data, we used the 6-item melancholia sub-scale (HDRS6) [23] instead of the complete HDRS17, consisting of the features shown on Table 5. Using the melancholia sub-scale is, therefore, twofold: it avoids the usage of the complete HDRS17 upon the available scarce dataset, whereas HDRS6 is able to capture the core symptoms of depression [23]. In addition, two psychotic symptoms were considered (hallucinations and delusions), totalizing eight features.

The *somatic general* item takes values on $\{0, 1, 2\}$, where the value 0 means the item is *absent*, and the value 2 means it is *clearly present*. The other items of HDRS6 are graded on $\{0, 1, 2, 3, 4\}$, where 0 means the item is *absent*, and 4 means the item is *severe* [22]. To use as much data as possible, the incomplete cases were imputed with the same method as in the original study [33], namely, the last observation carried forward (LOCF). The frequencies of the imputed data at each week are shown on Table 6. An additional step in data preprocessing to cope with the limitation of dataset size consisted of discretizing each item as binary variables on $\{low, high\}$, as follows: $\{0, 1\}$ was mapped to *low*, while $\{2, 3, 4\}$ (for five-valued variables) and $\{2\}$ (for the three-valued variable) were mapped to *high*.

Table 5
Summary of psychiatry data.

Psychotic Depression dataset [33]	
Number of sequences (complete)	122 (100) patients
Number of time points	8 (incl. baseline)
Depressive symptoms (HDRS6)	Depressed mood (Dm), Guilt, Work and Activities (Ac), Psychomotor Retardation (Re), Psychic Anxiety (Ap), and Somatic General (Sg)
Psychotic symptoms	Hallucinations (Ha) and Delusions (De)
Study's period and location	2002–2007, The Netherlands

6.3. Heuristic learning

Applying the heuristic procedure over the data first yields a DBN, with mean log-likelihoods −351.18. In the first iteration of the heuristic refinement, it tries to find a model with two cuts that is fitter than the DBN, which in fact was possible, precisely a PDBN-2 with cuts $\{4, 7\}$ and fit of −345.53, as show on left side of Fig. 3. Although not expanded further, the model with cuts $\{6, 7\}$ was also fitted better than the DBN (mean equal to −350.31). Since the algorithm found an improvement over the current best solution (the DBN), it updates the best solution to the most fit PDBN-2 and continues the heuristic search, now over PDBNs-3. As the right plot of Fig. 3 shows, the search again could find an improved solution, precisely a PDBN-3 with an additional cut just before the last cut, leading to a new cut set $\{4, 6, 7\}$ and mean log-likelihood of −344.80. Consequently, a new iteration is began over PDBNs-4, however, no further improvement could be achieved this time since the most fit PDBN-4 had a mean of −362.61 (plot not shown), leading to the termination of the procedure. Hence, the model returned was a PDBN-3 with cuts at $\{4, 6, 7\}$.

A more detailed examination of the time partitioning of the resulting PDBN-3 can reveal insight on the underlying dynamics of the psychiatric treatment. In general lines, it suggests that the dynamics governing roughly the first half of the treatment's duration is distinguished from the remaining weeks. The second half of treatment is further dichotomized since the transition pattern to the last week is distinguished as well. Hypothesis can be devised from this structural partitioning, e.g. whether there are one or more symptoms that have stronger influence on the others in the first stage, and whether the last transition is distinguished due to a possible stabilization. Nonetheless, clinically relevant questions as these need a stronger assessment based on the graphical structure and distributions of each of the three components of the model, as covered in the next section.

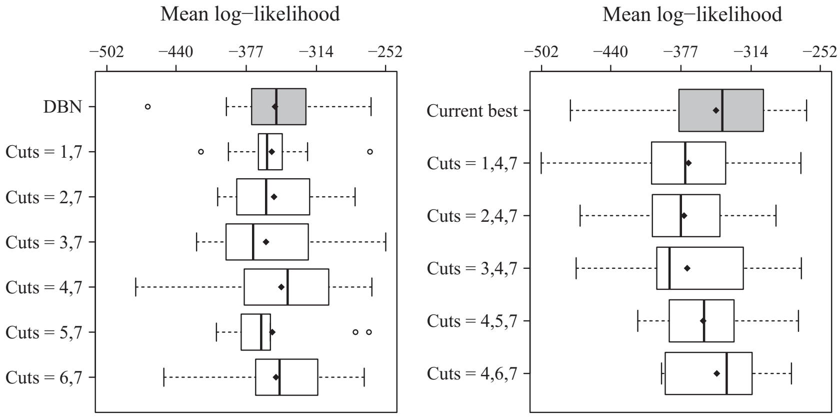
6.4. Transition structures

The structure of the DBN is shown on Fig. 4, while the structure of the conditional BNs that compose the PDBN-3 are shown on Figs. 5 and 6. For a clearer exposition, each conditional BN was split

Table 6
Relative frequencies of HDRS6 items of psychiatry data at each week, where *w* denotes the respective weighted means (in bold).

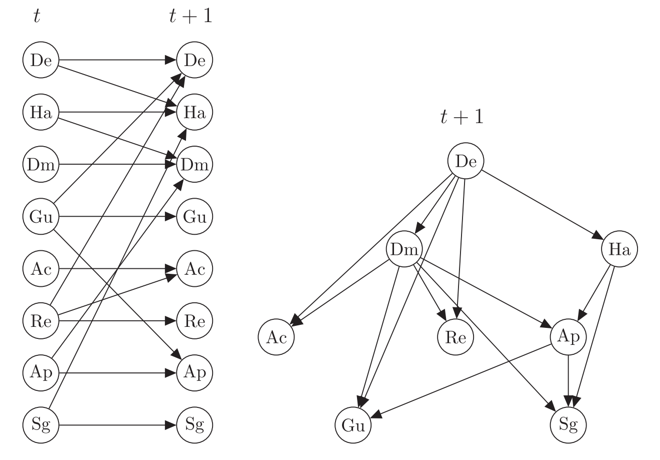
t	Depressed mood						Guilt						Work and activities					
	0	1	2	3	4	*w*	0	1	2	3	4	*w*	0	1	2	3	4	*w*
0	0	0	0.04	0.35	0.61	**3.57**	0.04	0.05	0.14	0.14	0.63	**3.27**	0	0	0.15	0.49	0.36	**3.21**
1	0.01	0.02	0.14	0.41	0.43	**3.23**	0.04	0.07	0.2	0.23	0.45	**2.98**	0	0	0.21	0.52	0.27	**3.06**
2	0.05	0.07	0.26	0.39	0.23	**2.69**	0.09	0.15	0.24	0.2	0.33	**2.52**	0	0.02	0.34	0.5	0.14	**2.76**
3	0.1	0.13	0.26	0.29	0.22	**2.4**	0.15	0.23	0.25	0.16	0.22	**2.07**	0.01	0.08	0.35	0.4	0.16	**2.61**
4	0.16	0.17	0.3	0.2	0.17	**2.07**	0.24	0.23	0.2	0.14	0.19	**1.81**	0.02	0.12	0.4	0.34	0.12	**2.43**
5	0.22	0.16	0.23	0.22	0.17	**1.97**	0.3	0.2	0.2	0.12	0.17	**1.67**	0.02	0.14	0.43	0.29	0.12	**2.34**
6	0.25	0.12	0.27	0.2	0.15	**1.87**	0.34	0.16	0.18	0.15	0.17	**1.66**	0.03	0.19	0.36	0.3	0.12	**2.29**
7	0.26	0.15	0.26	0.2	0.13	**1.79**	0.34	0.23	0.16	0.1	0.17	**1.52**	0.07	0.25	0.37	0.2	0.11	**2.03**

t	Psychomotor retardation						Psychic anxiety						Somatic general			
	0	1	2	3	4	*w*	0	1	2	3	4	*w*	0	1	2	*w*
0	0.16	0.3	0.31	0.22	0.02	**1.65**	0.03	0.14	0.27	0.37	0.19	**2.54**	0.1	0.3	0.61	**2.54**
1	0.15	0.33	0.34	0.16	0.02	**1.59**	0.11	0.16	0.29	0.29	0.16	**2.22**	0.16	0.34	0.51	**2.22**
2	0.27	0.3	0.29	0.12	0.02	**1.34**	0.18	0.22	0.3	0.23	0.07	**1.8**	0.22	0.43	0.34	**1.8**
3	0.33	0.35	0.22	0.08	0.02	**1.11**	0.29	0.25	0.23	0.16	0.07	**1.47**	0.34	0.39	0.27	**1.47**
4	0.4	0.31	0.2	0.07	0.02	**0.98**	0.3	0.26	0.2	0.17	0.06	**1.42**	0.27	0.48	0.25	**1.42**
5	0.53	0.21	0.18	0.06	0.02	**0.81**	0.39	0.2	0.24	0.12	0.05	**1.24**	0.39	0.42	0.2	**1.24**
6	0.52	0.27	0.13	0.06	0.02	**0.77**	0.39	0.16	0.23	0.17	0.04	**1.3**	0.41	0.38	0.21	**1.3**
7	0.62	0.18	0.12	0.06	0.02	**0.66**	0.38	0.26	0.19	0.12	0.05	**1.2**	0.4	0.36	0.24	**1.2**



(a) First iteration: 2 cuts. (b) Second iteration: 3 cuts.

Fig. 3. Boxplots for each stage of the heuristic over psychiatry data. The means are represented by a diamond symbol.



(a) Inter-temporal arcs. (b) Intra-temporal arcs.

Fig. 4. Structure of the DBN learned from the psychiatry data. Nodes on the left side of the inter-temporal arcs occur at time *t*, while those on the right at $t+1$. De = Delusions, Ha = Hallucinations, Dm = Depressed mood, Gu = Guilt, Ac = Work and activities, Re = Psychomotor retardation, Ap = Psychic anxiety, and Sg = Somatic general.

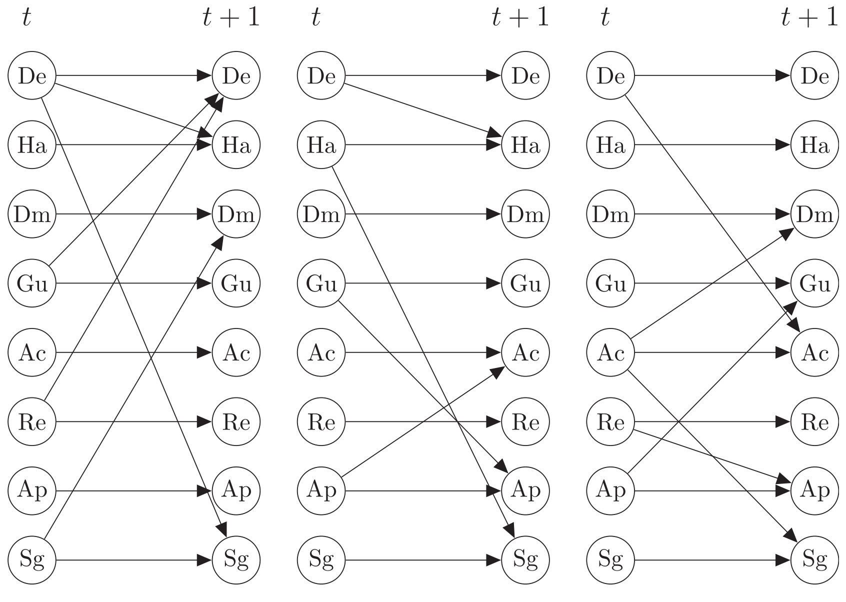
into *inter*-temporal arcs (i.e. those from $t+1$ to *t*) and *intra*-temporal arcs (those delimited to each point $t+1$). Note that DBN's and PDBN-3's initial structure are naturally the same. Both models indicate the existence of a self-influence for every feature when moving from present to future. More precisely, if *A* is a feature, the chain $A^{(t)} \rightarrow A^{(t+1)}$ has been regularly learned for both DBN and PDBN-3, indicating (part of) the direct effect received by $A^{(t+1)}$.

7. Model assessment from a clinical perspective

In this section we approach the use of the learned models for psychotic depression, specially the DBN and the PDBN-3, to support answering clinical-oriented questions.

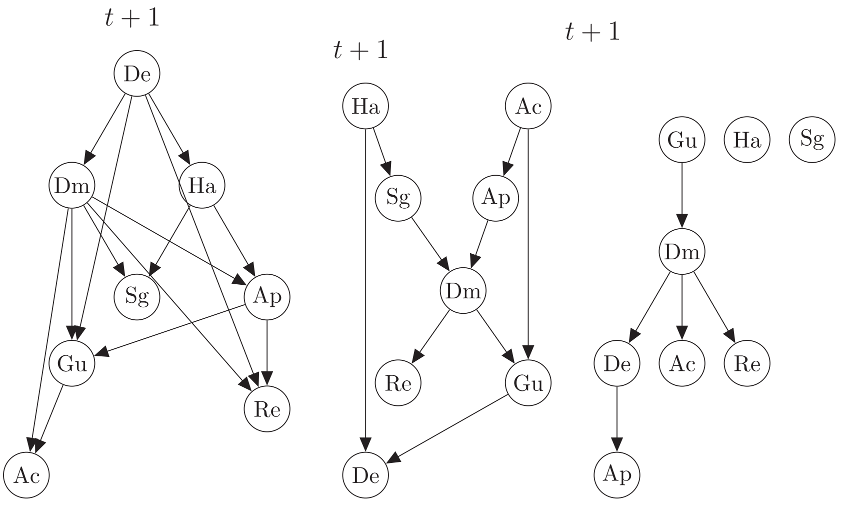
7.1. Symptoms' marginals over time

The previous sections showed that the PDBN-3 learned by the heuristic procedure provided: a better fit and a richer transition structure information with respect to other evaluated PDBNs, including the DBN. A complementary and practical assessment of these models compare the marginal frequencies of each symptom per week, as seen in data, with the respective model-based marginal distributions. Table 7 presents the empirical and model-based



(a) First cut: [0, 4]. (b) Second cut: [4, 6]. (c) Third cut: [6, 7].

Fig. 5. *Inter*-temporal arcs of the PDBN-3 learned from the psychiatry data.



(a) First cut: [0, 4]. (b) Second cut: [4, 6]. (c) Third cut: [6, 7].

Fig. 6. *Intra*-temporal arcs of the PDBN-3 learned from the psychiatry data.

marginals for each symptom per week, where the value assumed is either *true* or *high*. A summary of this information is presented at Table 8.

Concerning the psychotic symptoms, the PDBN-3 produced marginals that are closer to the empirical data than the DBN on average. With respect to depressive symptoms, a superior fit was achieved by the PDBN-3, except for the symptom psychomotor retardation.

7.2. Predictive symptoms over time

As discussed before, selecting an adequate structure is an important step to capture the underlying distribution in data as precisely as possible. As a probabilistic graphical model, the structure of PDBNs can be systematically verified for statistical independences among two sets of random variables by means of d-separation properties [8], essentially testing the paths between the respective nodes in the structure. As the Figs. 5 and 6 show, the marginal statistical dependences, both direct and indirect (i.e. through paths with two or more arcs), dominated over the marginal independences. Nevertheless, the independence relation $\perp\!\!\!\perp_P$ (or its counterpart $\not\perp\!\!\!\perp_P$) is qualitative, in the sense that two variables being dependent does not directly inform about any intensity in which this dependence occur.

In this context, we approach a research question within the field of psychiatry, specially in psychotic depression: *to which extent do psychotic and depressive symptoms interact during treatment*? This question can be rephrased more concretely as: *how predictive are the psychotic symptoms to depressive symptoms, and vice versa*? To answer this question, statistical (in)dependences play a key role, since it is the fundamental criterion to decide on dependence and independence. However, it must be complemented to allow

Table 7
Marginal distributions over time: psychiatry data and learned models (the latter minus the former). The time span is split according to the cut set of the PDBN-3. The values that are closer to the empirical frequencies are indicated in bold.

Symptom	Marginal probability (%)							
	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$
Delusions								
Data	91.0	72.1	59.0	47.5	40.2	36.1	32.0	30.3
DBN	−0.09	**0.43**	**0.32**	2.03	1.93	**0.22**	−**0.18**	−1.95
PDBN-3	−0.09	−0.88	−1.32	**0.28**	**0.16**	0.38	1.6	**0.11**
Hallucinations								
Data	23.8	15.6	16.4	13.1	13.1	11.5	13.9	11.5
DBN	0.03	3.69	−**0.25**	0.68	−**1.06**	−0.77	−4.26	−2.62
PDBN-3	0.03	**2.77**	−1.59	−**0.58**	−1.95	−**0.01**	−**2.05**	−**1.66**
Depressed mood								
Data	100.0	97.5	88.5	77.0	67.2	62.3	62.3	59.0
DBN	−0.83	−**4**	−**2.22**	2.02	4.96	4.07	−**1.07**	−2.06
PDBN-3	−0.83	−4.39	−3.66	−**1.08**	**0.67**	**4.02**	2.5	**1.76**
Guilt								
Data	91.0	88.5	76.2	62.3	53.3	50.0	50.0	42.6
DBN	−0.03	−**5.78**	−**2.37**	3.09	4.56	**1.49**	−3.76	−0.84
PDBN-3	−0.03	−6.72	−3.92	**0.9**	**2.03**	3	**1.17**	−**0.07**
Activities								
Data	100.0	100.0	98.4	91.0	86.1	83.6	77.9	68.0
DBN	−0.83	−4.36	−6.87	−3.87	−3.13	−4.52	−**2.36**	4.47
PDBN-3	−0.83	−**3.03**	−**4.14**	**0.16**	**1.73**	−**0.18**	2.72	**2.66**
Retardation								
Data	54.9	52.5	43.4	32.0	28.7	25.4	20.5	19.7
DBN	−0.1	−6.18	−4.38	**1.32**	−**0.01**	−**0.41**	1.77	**0.39**
PDBN-3	−0.1	−**4.3**	−**2.96**	1.78	−0.45	−2.73	−**0.86**	−2.32
Psychic anxiety								
Data	82.8	73.0	59.8	45.9	43.4	41.0	44.3	36.1
DBN	−0.01	−**4.76**	−**1.04**	5.93	3.04	**1.07**	−5.76	−**0.57**
PDBN-3	−0.01	−5.54	−3.19	**2.87**	−**0.56**	3.36	**0.17**	1.98
Somatic general								
Data	60.7	50.8	34.4	27.0	25.4	19.7	21.3	23.8
DBN	−0.02	−**6.71**	**0.83**	3.03	1.2	4.47	**0.94**	−**3.05**
PDBN-3	−0.02	−7.28	−0.95	**1.15**	−**0.29**	**1.57**	−1.63	−3.33

Table 8
Summary of percentage differences of learned models to the marginal frequencies of psychiatry data. The absolute values are used to compute the means. The best means are indicated in bold followed by an asterisk.

Feature	Mean diff. (DBN)	Mean diff. (PDBN-3)
Delusions	0.89	**0.6**
Hallucinations	1.67	**1.33***
Depressed mood	2.65	**2.36***
Guilt	2.74	**2.23***
Activities	3.8	**1.93***
Retardation	**1.82***	1.94
Psychic anxiety	2.77	**2.21***
Somatic general	2.53	**2.03***

an assessment of the intensity of dependence among different dependent variables, aiming ultimately at discovering adequate predictors, i.e. features capable of performing an effective prediction of the interested symptoms. Intuitively, a symptom is a good predictor if each of its groups (i.e. its values) induces a different distribution on the predicted symptom; in other words, it should allow to reasonably distinguish the predicted symptom.

In this section, the odds ratio criterion is employed to determine the strength of predictors. A subset of time points was selected as conditioning points to observe a psychotic (resp. depressive) symptom and then compute the ORs of future time points for each depressive (resp. psychotic) symptom. Using multiple points allows to evaluate the dynamics of predictive capability as treatment progresses and more information become available. These conditioning points were selected to match approximately the cut points of the PDBN-3 learned heuristically, namely, $\{1, 4, 6\}$. The baseline point ($t = 0$) was discarded since it was a weak predictor for most of these predictions.

In order to compute an OR, suppose X is a psychotic symptom observed at some point (e.g. at $t = 1$), and Y is a depressive symptom that will be predicted at $t = i, i > 1$; therefore, $val(X) = \{true, false\}$ and $val(Y) = \{low, high\}$. Then, the odds ratio to predict Y given X is:

$$\mathrm{OR}(Y^{(i)}|X^{(1)}) = \frac{\mathrm{odds}(Y^{(i)} = high|X^{(1)} = true)}{\mathrm{odds}(Y^{(i)} = high|X^{(1)} = false)} = \frac{P(Y^{(i)} = high|X^{(1)} = true)/(1 - P(Y^{(i)} = high|X^{(1)} = true))}{P(Y^{(i)} = high|X^{(1)} = false)/(1 - P(Y^{(i)} = high|X^{(1)} = false))} \tag{7}$$

We fix that each depressive variable Y is predicted with level *high*, hence, the OR indicates the chances of having level *high* in the future according to each group of a psychotic symptom X. If $\mathrm{OR} > 1$, then it is more likely that the depressive symptom Y will have level *high* if the patient comes from the group with $X = true$ compared to the patients coming from the group $X = false$; if $\mathrm{OR} < 1$, it is more likely to observe Y at *high* in the group $X = false$ than in the group $X = true$; finally, if $\mathrm{OR} = 1$, there is no association between X and Y, i.e. knowing the group of this particular psychotic symptom does not affect the predictions for this depressive symptom. For the sake of terminology, an $\mathrm{OR} > 1$ is also called a positive correlation, while an $\mathrm{OR} < 1$ indicates a negative correlation. Note that for the case when X is depressive and Y is psychotic, we fix *true* for X, and *high* and *low* in the numerator and denominator for Y respectively.

Additionally, to evaluate of the significance of the association between each X and Y, tables of contingency were constructed based on expected counts from the model. The Fisher's exact test was employed to evaluate the statistical significance of these, under a significance level of $\alpha = 0.05$.

7.2.1. Predictors for depression

Table 9a shows the ORs for psychotic symptoms one week after baseline (i.e. at $t = 1$), acting as predictors for depression. These results suggest that delusions at that point had an at least reasonable association with the symptoms depressed mood and guilt, i.e. for at least half of the future points that were predicted. On the other hand, hallucinations at $t = 1$ showed to be less associated to the depressive symptoms. Nonetheless, somatic general contrasts with this pattern, as it has been predicted by hallucinations almost until the end of the remaining weeks of treatment. The other case where some dependency on this predictor was noticed is psychic anxiety, however for a shorter period of time (three weeks forward).

Table 9
Odds ratios for **psychotic symptoms as predictors**. An OR greater than 1 indicates that the level *high* on the depressive symptom is more likely to be observed in the group *true* than in the group *false* of the psychotic symptom. Results marked in bold and * stand for a statistically significant association.

Symptom and predictor	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$
(a) Odds ratios based on $t = 1$						
Depressed mood						
Delusions$^{(1)}$	**5.15***	**3.39***	**2.72***	1.75	1.38	1.44
Hallucinations$^{(1)}$	1.13	1.5	1.46	1.59	1.66	1.48
Guilt						
Delusions$^{(1)}$	**3.84***	**3.27***	**2.75***	**2.11***	1.84	1.62
Hallucinations$^{(1)}$	1.1	1.12	1.2	1.2	1.3	1.29
Activities						
Delusions$^{(1)}$	3.53	2.23	2.45	1.42	1.4	1.45
Hallucinations$^{(1)}$	1.34	1.04	1.38	1.38	1.6	1.47
Retardation						
Delusions$^{(1)}$	**3.24***	**3.22***	2.4	2.02	1.67	1.35
Hallucinations$^{(1)}$	1.15	1.16	1.24	1.33	1.25	1.35
Psychic anxiety						
Delusions$^{(1)}$	1.33	1.21	1.16	1.27	1.33	1.46
Hallucinations$^{(1)}$	**2.54***	**2.66***	**2.41***	1.65	1.32	1.31
Somatic general						
Delusions$^{(1)}$	0.96	0.95	0.8	0.7	0.64	0.82
Hallucinations$^{(1)}$	**3.31***	**3.27***	**2.86***	**3.07***	**2.97***	2.23

Symptom and predictor	$t = 5$	$t = 6$	$t = 7$	Symptom and predictor	$t = 7$
(b) Odds ratios based on $t = 4$ (left) and $t = 6$ (right)					
Depressed mood				Depressed mood	
Delusions$^{(4)}$	**3.09***	**2.26***	**2.17***	Delusions$^{(6)}$	**2.72***
Hallucinations$^{(4)}$	1.98	2.14	1.71	Hallucinations$^{(6)}$	1.67
Guilt				Guilt	
Delusions$^{(4)}$	**4.15***	**2.93***	**2.34***	Delusions$^{(6)}$	**3.62***
Hallucinations$^{(4)}$	1.19	1.31	1.4	Hallucinations$^{(6)}$	1.2
Activities				Activities	
Delusions$^{(4)}$	2.26	1.81	**2.59***	Delusions$^{(6)}$	**5.66***
Hallucinations$^{(4)}$	2.52	1.53	1.61	Hallucinations$^{(6)}$	1.61
Retardation				Retardation	
Delusions$^{(4)}$	**2.97***	2.02	1.98	Delusions$^{(6)}$	2.04
Hallucinations$^{(4)}$	1.4	1.25	1.36	Hallucinations$^{(6)}$	1.34
Psychic anxiety				Psychic anxiety	
Delusions$^{(4)}$	1.88	1.88	**2.21***	Delusions$^{(6)}$	**3.52***
Hallucinations$^{(4)}$	2.18	1.53	1.45	Hallucinations$^{(6)}$	1.25
Somatic general				Somatic general	
Delusions$^{(4)}$	0.97	0.87	0.99	Delusions$^{(6)}$	1.14
Hallucinations$^{(4)}$	**6.52***	**6.18***	**4.91***	Hallucinations$^{(6)}$	**4.31***

With respect to the predictive power of psychotic symptoms observed at $t = 4$ and $t = 6$ (Table 9b, left and right respectively), delusions stood as predictor of depressed mood and guilt, in this situation as a stronger predictor (all three future predictions were significant). Other depressive symptoms were mostly weakly associated to delusions. Hallucinations at these time points showed a more restricted behavior than before, since it acted as predictor of somatic general only, although by significant associations.

7.2.2. Predictors for psychosis

In the following, we evaluate how predictive the depressive symptoms are to predict psychotic symptoms. Note that ORs are not symmetric; for example, we calculate $P(\text{Som.gen}^{(t)}|\text{Del}^{(0)})$ to assess whether delusions is predictive to somatic general, while we compute $P(\text{Del}^{(t)}|\text{Som. gen}^{(0)})$ to assess whether somatic general is predictive to delusions. Note that these two might represent distinct quantities.

Table 10a shows the odds ratio for each depressive symptom observed at $t = 1$. As the results indicate, the depressive symptoms were not significantly strong to predict delusions, except depressed mood, guilt and retardation, which accounted for a weak association (precisely, two weeks ahead of the reference measurement). Regarding hallucinations, there is virtually no depressive symptom predictor for the case of $t = 1$.

Table 10
Odds ratios for **depressive symptoms as predictors**. An OR greater than 1 indicates that the level *true* on the psychotic symptom is more likely to be observed in the group *high* than in the group *low* of the depressive symptom. Results marked in bold and * stand for a statistically significant association.

Symptom and predictor	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$
Odds ratios based on $t = 1$						
Delusions						
Depressed mood$^{(1)}$	**5.3***	**6.91***	5.04	4.2	3.66	3.21
Guilt$^{(1)}$	**2.91***	**2.86***	2.21	2.16	1.9	1.62
Activities$^{(1)}$	2.79	2.78	2.05	1.75	1.53	1.31
Retardation$^{(1)}$	**2.49***	**2.11***	1.8	1.57	1.37	1.38
Psychic anxiety$^{(1)}$	1.18	1.19	1.18	1.26	1.27	1.22
Somatic general$^{(1)}$	0.91	0.97	0.96	1.07	1.07	1.16
Hallucinations						
Depressed mood$^{(1)}$	0.58	0.49	0.42	0.83	0.9	0.75
Guilt$^{(1)}$	0.84	0.86	0.78	0.78	0.79	0.67
Activities$^{(1)}$	0.51	0.44	0.38	0.38	0.41	0.31
Retardation$^{(1)}$	1.08	0.93	0.91	0.81	0.93	1.07
Psychic anxiety$^{(1)}$	1.84	2.05	1.71	1.83	1.39	1.52
Somatic general$^{(1)}$	**3.04***	2.9	2.54	1.86	1.86	1.94

Symptom and predictor	$t = 5$	$t = 6$	$t = 7$	Symptom and predictor	$t = 7$
Odds ratios based on $t = 4$ (left) and $t = 6$ (right)					
Delusions				Delusions	
Depressed mood$^{(4)}$	**4.96***	**3.97***	**4.22***	Depressed mood$^{(6)}$	**3.94***
Guilt$^{(4)}$	**8.13***	**5.62***	**4.58***	Guilt$^{(6)}$	**5.63***
Activities$^{(4)}$	3.84	3.36	3.14	Activities$^{(6)}$	1.83
Retardation$^{(4)}$	**3.32***	**2.5***	**2.2***	Retardation$^{(6)}$	1.87
Psychic anxiety$^{(4)}$	1.8	1.69	1.9	Psychic anxiety$^{(6)}$	**2.52***
Somatic general$^{(4)}$	1.35	1.35	1.37	Somatic general$^{(6)}$	1.19
Hallucinations				Hallucinations	
Depressed mood$^{(4)}$	1.2	1.2	0.97	Depressed mood$^{(6)}$	1.71
Guilt$^{(4)}$	1.07	0.91	0.96	Guilt$^{(6)}$	1.35
Activities$^{(4)}$	0.82	0.9	0.67	Activities$^{(6)}$	1.25
Retardation$^{(4)}$	1.04	1.3	1.27	Retardation$^{(6)}$	1.41
Psychic anxiety$^{(4)}$	**3.8***	**3***	2.97	Psychic anxiety$^{(6)}$	1.29
Somatic general$^{(4)}$	**4.85***	**3.6***	**3.36***	Somatic general$^{(6)}$	**7.47***

On the other hand, updating the depressive symptoms at $t = 4$, as shown on Table 10b (left), increased the association of the three symptoms mentioned before to predict delusions until the end. The same insight applies to predict delusions at $t = 6$. Concerning the prediction of hallucinations, somatic general emerged with strong associations when measured both at $t = 4$ and $t = 6$, while psychic anxiety showed reasonable associations only when measured at the middle point, though.

8. Conclusions

In this work, we proposed a heuristic algorithm to learn non-homogeneous time dynamic Bayesian networks for relatively small temporal datasets with a small number of variables as typically encountered in clinical settings. Extensive simulations and a case study in psychiatry (psychotic depression) demonstrated its capability to find adequate models under different assumptions, which included data generated from non-homogeneous and homogeneous models. In particular, simulating experiments played an important role to show that, in more general scenarios, models based on non-homogeneous time have substantial benefits over DBNs on several dimensions (e.g. model fit and problem insight) when the underlying process switches between different regimes over time. In the case of small datasets, which are commonly found in many clinical studies, the results indicate that the heuristic algorithm behaves in a more conservative fashion, i.e. it tends to produce slightly simpler non-homogeneous models compared to the reference models, and yet providing a decent fit.

Aiming at learning non-homogeneous models in the usually unfavorable scenario of data scarcity, an evaluation criterion was employed by the heuristic to explicitly avoid over-specialized models, taking into account the need for robustness. Moreover, the empirical results suggest that the search strategy of the heuristic, which is based on an incremental construction of non-homogeneous models, is able to properly cope with the trade-off between model complexity and data scarcity.

A first step towards a systematic application of probabilistic graphical models in psychiatry taking into account the temporal dimension was taken. It allowed to obtain insight about the dynamics of this medical condition over the duration of a controlled treatment. In particular, a research question aiming to answer the temporal relationship between psychotic and depressive symptoms was investigated, supported by models learned with the heuristic procedure. The experimental assessment of the predictive capability of psychotic symptoms observed at different moments (near baseline, middle and near-end points) showed that the delusions symptom was more predictive than the hallucinations symptom on most cases. On the other hand, the depressive symptoms were less predictive for the psychotic symptoms. Nevertheless, a point to be observed it that in general the predictions were bidirectional, i.e. the symptoms from one category that stood as statistically significant predictors for the other can be interchanged.

Among future research, we intend to evaluate the developed algorithm on other real-world problems, as well as investigate further variations of the incremental search. For example, during the execution of the algorithm, different new solutions with equal or approximately equal score yet higher than the current best solution can be found; this is currently worked out choosing one of these new solutions randomly and then continuing the search. The problem of handling multiple solutions is in fact recurring in the literature of Bayesian networks, where extensive research has been developed [34–36,20]. In this direction, the approach of this paper could benefit from such research, for example by extending the greedy search, as well as taking into account Bayesian approaches [15]. These further investigations could provide more insight about the distribution and the variance of the cut sets.

Acknowledgments

The authors would like to thank J. Wijkstra and W.A. Nolen for providing the psychiatry data used in this work.

This work has been funded by NWO (Netherlands Organisation for Scientific Research).

The original Psychotic Depression study (DUDG) was supported by grants from AstraZeneca and Wyeth Pharmaceuticals.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.05.003.

References

[1] C. Rose, C. Smaili, F. Charpillet, A dynamic Bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis, in: 17th IEEE International Conference on Tools with Artificial Intelligence, 2005, ICTAI 05, 2005, p. 5. pp. 598.

[2] M.A. van Gerven, B.G. Taal, P.J. Lucas, Dynamic Bayesian networks as prognostic models for clinical patient management, J. Biomed. Inf. 41 (4) (2008) 515–529.

[3] D. Kim, J. Burge, T. Lane, G. Pearlson, K. Kiehl, V. Calhoun, Hybrid ICA-Bayesian network approach reveals distinct effective connectivity differences in schizophrenia, NeuroImage 42 (4) (2008) 1560–1568.

[4] J. Burge, T. Lane, H. Link, S. Qiu, V.P. Clark, Discrete dynamic Bayesian network analysis of fMRI data, Hum. Brain Map. 30 (1) (2009) 122–137.

[5] Z. Shojaei Estabragh, M. Riahi Kashani, F. Jeddi Moghaddam, S. Sari, Z. Taherifar, S. Moradi Moosavy, K. Sadeghi Oskooyee, Bayesian network modeling for diagnosis of social anxiety using some cognitive-behavioral factors, Netw. Model. Anal. Health Inf. Bioinf. 2 (4) (2013) 257–265.

[6] M.H. DeGroot, M.J. Schervish, Probability and Statistics, third ed., Pearson, 2011.

[7] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[8] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, The MIT Press, 2009.

[9] K. Orphanou, A. Stassopoulou, E. Keravnou, Temporal abstraction and temporal Bayesian networks in clinical domains: a survey, Artif. Intell. Med. 60 (3) (2014) 133–149.

[10] T. Charitos, L.C. van der Gaag, S. Visscher, K.A. Schurink, P.J. Lucas, A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients, Expert Syst. Appl. 36 (2, Part 1) (2009) 1249–1258.

[11] F.L. Seixas, B. Zadrozny, J. Laks, A. Conci, D.C.M. Saade, A Bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment, Comp. Biol. Med. 51 (2014) 140–158.

[12] M. van der Heijden, M. Velikova, P.J. Lucas, Learning Bayesian networks for clinical time series analysis, J. Biomed. Inf. 48 (2014) 94–105.

[13] F. Dondelinger, S. Lèbre, D. Husmeier, Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure, Mach. Learn. 90 (2) (2013) 191–230.

[14] S. Lèbre, J. Becq, F. Devaux, M. Stumpf, G. Lelandais, Statistical inference of the time-varying structure of gene-regulation networks, BMC Syst. Biol. 4 (1) (2010) 1–16.

[15] J.W. Robinson, A.J. Hartemink, Learning non-stationary dynamic Bayesian networks, J. Mach. Learn. Res. 11 (2010) 3647–3680.

[16] M. Grzegorczyk, D. Husmeier, Non-stationary continuous dynamic Bayesian networks, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems, vol. 22, 2009, pp. 682–690.

[17] T. Charitos, L.C. van der Gaag, Sensitivity analysis for threshold decision making with dynamic networks, in: UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13–16, 2006.

[18] S. Visscher, P. Lucas, I. Flesch, K. Schurink, Using temporal context-specific independence information in the exploratory analysis of disease processes, in: R. Bellazzi et al. (Eds.), Artificial Intelligence in Medicine, Lecture Notes in Computer Science, vol. 4594, Springer, Berlin H., 2007, pp. 87–96.

[19] A. Tucker, X. Liu, Learning dynamic Bayesian networks from multivariate time series with changing dependencies, in: M.R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, C. Borgelt (Eds.), Advances in Intelligent Data Analysis V, Lecture Notes in Computer Science, vol. 2810, Springer, Berlin Heidelberg, 2003, pp. 100–110.

[20] K. Murphy, Dynamic Bayesian Networks: Representation, Inference and Learning, Ph.D. thesis, UC Berkeley, Computer Science Division, July 2002.

[21] J. Wijkstra, Pharmacological Treatment of Psychotic Depression: In Search for Evidence Ph.D. thesis, Rijksuniversiteit Groningen, 2010.

[22] M. Hamilton, A rating scale for depression, J. Neurol., Neurosurg., Psych. 23 (1) (1960) 56.

[23] C.L. Hooper, D. Bakish, An examination of the sensitivity of the six-item Hamilton rating scale for depression in a sample of patients suffering from major depressive disorder, J. Psych. Neurosci. 25 (2) (2000) 178–184.

[24] W. Zucchini, An introduction to model selection, J. Math. Psychol. 44 (1) (2000) 41–61.

[25] T.M. Cover, J.A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing), Wiley-Interscience, 2006.

[26] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, 2001.

[27] G. Melançon, F. Philippe, Generating connected acyclic digraphs uniformly at random, Inf. Process. Lett. 90 (4) (2004) 209–213.

[28] J.-P. Chevrolat, J.-L. Golmard, S. Ammar, R. Jouvent, J.-F. Boisvieux, Modelling behavioral syndromes using Bayesian networks, Artif. Intell. Med. 14 (1998) 259–277.

[29] M. Klein, G. Modena, Estimating mental states of a depressed person with Bayesian networks, in: M. Ali, T. Bosse, K.V. Hindriks, M. Hoogendoorn, C.M. Jonker, J. Treur (Eds.), Contemporary Challenges and Solutions in Applied Artificial Intelligence, Studies in Computational Intelligence, vol. 489, Springer International Publishing, 2013, pp. 163–168.

[30] D.-I. Curiac, G. Vasile, O. Banias, C. Volosencu, A. Albu, Bayesian network model for diagnosis of psychiatric diseases, in: Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, 2009, ITI '09, 2009, pp. 61–66.

[31] Y.-S. Chang, C.-T. Fan, W.-T. Lo, W.-C. Hung, S.-M. Yuan, Mobile cloud-based depression diagnosis using an ontology and a Bayesian network, Fut. Gener. Comp. Syst. 43–44 (2015) 87–98.

[32] S. Sorias, Bayesian networks: overcoming the limitations of the descriptive and categorical approaches in psychiatric diagnosis. a proposal based on Bayesian networks, Turk. J. Psych. (2014) 1–12.

[33] J. Wijkstra, H. Burger, W.W. Van Den Broek, T.K. Birkenhäger, J.G.E. Janzing, M. P.M. Boks, J.A. Bruijn, M.L.M. Van Der Loos, L.M.T. Breteler, G.M.G.I. Ramaekers, R.J. Verkes, W.A. Nolen, Treatment of unipolar psychotic depression: a randomized, double-blind study comparing imipramine, venlafaxine, and venlafaxine plus quetiapine, Acta Psych. Scand. 121 (3) (2009) 190–200.

[34] J. Cheng, R. Greiner, J. Kelly, D. Bell, W. Liu, Learning bayesian networks from data: an information-theory based approach, Artif. Intell. 137 (2002) 43–90.

[35] R. Castelo, T. Kocka, On inclusion-driven learning of Bayesian networks, J. Mach. Learn. Res. 4 (2003) 527–574.

[36] P. Larrañaga, M. Poza, Y. Yurramendi, R.H. Murga, C.M.H. Kuijpers, Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters, IEEE Trans. Pattern Anal. Mach. Intell. 18 (9) (1996) 912–926.