Unraveling temporal processes using probabilistic graphical models

de Paula Bueno, M.L.

6

PARTITIONED DYNAMIC BAYESIAN NETWORKS

When modeling the dynamics of real-world processes, the model properties are often assumed to be constant over time, resulting in a so-called time-homogeneous process. This might be justified, e.g., by scarce amounts of data available. While this reduces the number of parameters to be learned from data, the specificities of the underlying process are to some degree lost in the obtained models. In this chapter, we propose partitioned dynamic Bayesian networks *for capturing distribution regime changes, 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 such models taking into account the preference for less complex models. Experiments are performed based on simulated data to evaluate how well the proposed method is able to recover the original distributions, for different assumptions regarding the data generating mechanism. Finally, we consider a study case based on psychotic depression complementary to that of Chapter 4 to evaluate the goodness-of-fit and insight that partitioned dynamic Bayesian networks can provide to a real-world problem.*

6.1 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.

One 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 [29,

74, 101, 146, 157]. Therefore, the model has the same properties for every time point, as modeled by the well-known first-order homogeneous Markov chains [52]. A generalization of Markov chains to multivariate problems are the dynamic Bayesian networks [104, 136], which have been applied to a number of real-world domains, such as medicine [38, 86, 132, 153] and bioinformatics [54, 109, 145]. 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 DBNs solve the robustness problem, they introduce an undesirable effect: there is no distribution specificity as a function of time. Hence, one will never learn the details of the underlying process as was the aim in the first place.

It is known that in many clinical situations the dependences between symptoms and signs might change over time, as in the case of intervention studies where different sets of correlations are expected to occur 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 potentially better model fit and more insight that really helps in understanding the underlying process.

Although the notion of non-homogeneous models (a shorthand for *non-homogeneous time models*) is certainly not new, it is often the case that such models employ a number of approximations, for example due to properties of the targeted applications. Typically, non-homogeneous PGMs have been focused on biological processes, where regime shifts are assumed to be smooth [79, 109, 145]. These assumptions might, however, not be natural for other processes, 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 cut-off points to obtain new specific models, 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 chapter, we first introduce *partitioned dynamic Bayesian networks* (PDBNs, for short), which allow to express a process as a collection of DBNs. PDBNs make few assumptions regarding the process, the main one being the fact that the process duration is partitioned in the same way for every observable variable involved. Then, we propose a heuristic procedure to explore the space of PDBNs, taking into account the balance between specificity and simplicity. The approach starts with a 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 model fit over a training and test setting.

In order to demonstrate the applicability of the proposed model and heuristic method, an extensive set of simulations and real-world-based experiments are carried out. In simulations we evaluate whether the heuristic algorithm is able to recover adequate models in terms of statistical distance to the data generating

model, be it a homogeneous or a non-homogeneous model. We also aim to evaluate experimentally the behavior of the heuristic in the case of small datasets. Additionally, we consider a study case on psychotic depression data, and evaluate the homogeneous and non-homogeneous models learned from this data. Based on the obtained models, research questions of clinical relevance are formulated regarding the prediction of symptom association over time.

The remainder of this chapter is organized as follows. Section 6.2 describes related literature on homogeneous and non-homogeneous dynamic Bayesian networks in clinical and biological domains. Partitioned DBNs and the heuristic procedure to learn PDBNs are presented in Section 6.3. Simulations to evaluate the learning procedure are discussed in Section 6.4, while the models learned from psychiatry data are discussed in Section 6.5. Clinically-oriented discussions based on the psychiatry models are provided in Section 6.6, and lastly Section 6.7 gives the conclusions and suggestions for future research.

6.2 RELATED WORK

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 DBNs has been explored for finding correlations among different brain regions in several disorders, such as schizophrenia [101] and Alzheimer's disease [29]. 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 pneumonia [37].

Another aspect of DBNs explored in the clinical domain is the predictive ability for several tasks, e.g. diagnosis [38, 146] and prognosis [74]. 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 evolves. 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 [38]. The application of DBNs and similar models in clinical domains has been compared to similar formalisms in a recent survey [132].

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 [171]. On the other hand, more sophisticated models have been developed in other fields, mainly biological processes [54, 79, 109, 145]. These models are constructed based on assumptions justified by domain knowledge;

for example, in some biological processes the intensity of interactions change over time, but no interaction is created or destroyed [79].

The aforementioned non-homogeneous models assume a set of assumptions or use a specific learning methodology, which we summarize as follows. Firstly, additional restrictions are usually imposed to the model structure, ranging from constrained intra-temporal interactions [109, 145] to completely fixed structure with flexibility on the parameter space only [79]. A second assumption is that regime switch in the process occurs in a smooth fashion. Finally, in many biological-oriented networks the learning approach is based on sampling strategies [54, 79, 109, 145], which can depend on additional assumptions in order to be feasible. As we show further in the chapter, these assumptions will not be considered for the development of PDBNs. Other approaches include, e.g., DBN models with hidden variables to control the dependence structure, which has been applied to engineering problems [170].

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. Imposing the previous assumptions on the manner by 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.

6.3 PARTITIONED DYNAMIC BAYESIAN NETWORKS

Models of non-homogeneous time can be defined by a set of transition distributions that should hold at specific intervals of the considered time series. In this work, the central idea lies in 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. 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.

6.3.1 *Model specification*

Definition 6.1 (Time partition)**.** *A time partition of a set of integers $\{0, \ldots, T\}$ is a set of integers $\{t_1, \ldots, t_k\}$, where $t_1 > 0$, $t_k = T$, and $t_i < t_{i+1}$ for $1 \leq i < k$. Each t_i, with $i > 1$, defines a set $\{1 + t_{i-1}, \ldots, t_i\}$, and t_1 defines the set $\{0, \ldots, t_1\}$.*

We say that each element of the time partition is a cut (a shorthand for cut-off) and we say that such time partition has k cuts.

The aim of Definition 6.1 is to split a time series horizon into a partition of indices. For example, given a time series indexed by the time points $\{0, \ldots, 7\}$, the time partition $\{2, 7\}$ has 2 cuts and splits the time series as follows: $\{0, 1, 2\}$,

and $\{3,4,5,6,7\}$. This definition is useful for defining non-homogeneous models as follows.

Definition 6.2 (Partitioned dynamic Bayesian network)**.** *Consider a time partition with k cuts of the integers $\{0,\ldots,T\}$, where the ith cut is associated to a conditional Bayesian network $\mathcal{B}_i$ over $\mathbf{X}^{(t+1)}$ conditioned on $\mathbf{X}^{(t)}$, $t \geq 0$. A partitioned dynamic Bayesian network with k cuts, denoted by PDBN-k, is a dynamic system $(\mathcal{B}_0,\ldots,\mathcal{B}_k)$ over $\mathbf{X}$ where:*

- *$\mathcal{B}_0 = (\mathcal{G}_0, P_0)$ is a Bayesian network over the variables $\mathbf{X}^{(0)}$ called initial network.*
- *$\mathcal{B}_i = (\mathcal{G}_i, P_i)$, $i > 0$, is a conditional Bayesian network over the variables $\{\mathbf{X}^{(t)}, \mathbf{X}^{(t+1)}\}$ called the ith transition network. The transition model $\mathcal{B}_i$ is associated to the ith cut of the time partition.*

We use the term *distribution cut* to denote a cut in the context of a PDBN. The joint distribution of an unrolled PDBN can be obtained by unrolling the transition models over the time points each transition model is associated to. This is as follows: the structure and parameters of all the nodes at time $t = 0$ come from the initial model $\mathcal{B}_0$, while the structure and parameters for any node $X_i^{(t)}$, where $t > 0$, come from the transition model whose cut includes t, i.e., the B_i such that $t \in \{1 + t_{i-1},\ldots,t_i\}$. Therefore, the joint distribution of an unrolled PDBN with k cuts $\{t_1,\ldots,t_k\}$ is as follows:

$$\begin{aligned} P(\mathbf{X}^{(0:T)}) = &\prod_{i=1}^{n} P_0(X_i^{(0)} \mid \pi(X_i^{(0)}, \mathcal{B}_0)) \\ &\cdot \prod_{r=1}^{k} \prod_{t=t_{r-1}}^{t_r-1} \prod_{i=1}^{n} P_r(X_i^{(t+1)} \mid \pi(X_i^{(t+1)}, \mathcal{B}_r)) \end{aligned} \tag{6.1}$$

where $t_0 = 0$ and P_r refers to the CPTs pertaining to the transition model $\mathcal{B}_r$. Note that the parent set of each X_i depends on $\mathcal{B}_r$ as denoted by $\pi(X_i, \mathcal{B}_r)$.

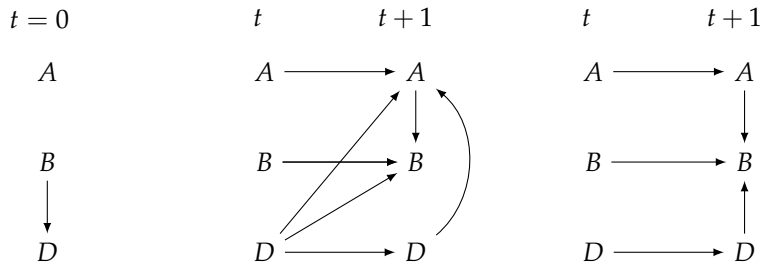
It follows from the previous definitions that a DBN is a PDBN with a single cut $\{T\}$, hence, a DBN is a PDBN-1.

Example 6.1. *Consider again the situation of Example 2.2, where two symptoms A and B and a drug quantity D are measured per patient on a regular basis. We define a PDBN-2 for this problem consisting of two cuts $\{2,7\}$ whose initial structure and transition structures are shown on Fig. 6.1. Each cut of the PDBN is associated to a conditional BN as follows: $\mathcal{B}_1$ holds for the time points $\{0,1,2\}$, while $\mathcal{B}_2$ holds for the time points $\{3,4,5,6,7\}$.*

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

$$\begin{aligned} P(\mathbf{X}^{(0:7)}) = &\prod_{i} P_0(X_i^{(0)} \mid \pi(X_i^{(0)}, \mathcal{B}_0)) \\ &\cdot \prod_{0 \leq t \leq 1} \prod_{i} P_1(X_i^{(t+1)} \mid \pi(X_i^{(t+1)}, \mathcal{B}_1)) \\ &\cdot \prod_{2 \leq t \leq 6} \prod_{i} P_2(X_i^{(t+1)} \mid \pi(X_i^{(t+1)}, \mathcal{B}_2)) \end{aligned} \tag{6.2}$$

where $X_i \in \mathbf{X}$, $\mathbf{X} = \{A, B, D\}$, *and* P_i *refers to the CPTs pertaining to the transition model* $\mathcal{B}_i$.



(a) Initial model $\mathcal{B}_0$ (b) Transition model $\mathcal{B}_1$ (c) Transition model $\mathcal{B}_2$

Figure 6.1: An example of PDBN-2. $\mathcal{B}_i$ represents the ith transition model (only its structure is shown, parameters are omitted). Nodes on the left and right side occur at t and $t+1$ respectively, except for the initial model.

6.3.2 *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 beneficial, 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.

6.3.2.1 *Algorithm description*

Taking the aforementioned factors into account, we present 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 splitting one of the transition distributions of the current PDBN. At each iteration a new cutting point is added without eliminating the cuts previously found. The procedure is greedy since it does not further explore the branching of solutions that are less interesting at each iteration. It is important to consider a strategy with feasible running time to search over the space of PDBNs, since the number of possible manners in which a discrete time series can be partitioned is potentially large. 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 2. 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 homogeneous model. Let us denote by s the size of the current cut set, i.e., $s = |C|$. By entering the outer loop (Line 2) the algorithm 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 5-6). 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 is better than 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 $k' \leq k$.

Algorithm 2 Builds a PDBN

Input: D: a dataset of sequences with length $\{0, \ldots, T\}$;
k: the maximum size of the cut set, $1 \leq k \leq T$.
Output: a PDBN-k', where $k' \leq k$.

1: $C \leftarrow \{T\}$
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 $\mathbf{C} = \{C_1, \ldots, C_r\}$.
4: Evaluate each cut set in $\mathbf{C}$ by means of a criterion f.
5: **if** there is a new cut set $C_i \in \mathbf{C}$, where $1 \leq i \leq r$, 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** PDBN-k' with cut set C learned from the data D.

6.3.2.2 *Evaluation criterion*

As Algorithm 2 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 [182]. 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 model likelihood) and information theory criteria (e.g. the Akaike information criterion and the Bayesian information criterion) [46].

For example, in order to employ the AIC in Algorithm 2, one would first learn a PDBN from the full dataset (i.e. all the sequences, hence a DBN) using the

AIC as scoring function. Then, each sub-DBN associated to new cut sets (Line 3) would be learned based on this score using the corresponding part of the data.

6.3.2.3 *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.

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. Note that this construction is sound since the model is time-homogeneous. If the original dataset D consists of m sequences (each of length $T + 1$), this merged dataset will consist of mT short sequences (each of length 2). Thus, abstracting the cost of learning a DBN by means of a cost function g will lead to a cost of $\mathcal{O}(g(mT))$ for learning 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 sequences, each with length 2. As each of these sub-DBNs would cost $g(m)$, learning such PDBN would require $\mathcal{O}(Tg(m)))$.

6.4 EMPIRICAL EVALUATION VIA SIMULATIONS

6.4.1 *Simulation parameters*

In this section experiments based on simulated data are presented for a general assessment of the proposed method for learning PDBNs. Time series with varying length and number of sequences were generated, resulting in diversified datasets. We considered the number of features as $n \in \{2, 6, 10, 14, 18\}$, and defined that each time series is composed by sequences with length of 10 or 30 time points. Hence, the unrolled models used in simulations have between 20 and 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, the simulation cases allow for a reasonable evaluation in terms of different feature spaces and dataset sizes.

For each simulation scenario, 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 the graphical structure of each random sub-DBN was uniformly generated at random [122], and distribution parameters

determined randomly as well (no noise was introduced in the model's parameters). Hence, each node of an unrolled PDBN assumes 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 distinct 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 aims at studying the effect on the heuristic's capabilities over different quantities of data.

Each dataset was generated from either a random DBN or a random PDBN. The initial aim is to verify experimentally whether the construction algorithm is able to learn the adequate class of model with respect to the reference model (a random DBN or PDBN) used to simulate data. Moreover, the cuts of the learned models are 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 model, we write '$\subseteq +a$', where a denotes the number of additional cuts included by the learned model.
- If none of these criteria is met, we write '$\not\subseteq$'.

Although this notation is useful to perform a structural comparison in terms of the number and position of distribution cuts, they do not provide information about the distance between the probability distributions of two models. To this end, the Kullback-Leibler (KL, for short) divergence [46] between the marginal distribution of each feature $X_i^{(t)}$ was considered, which indicates the amount of additional information one needs to codify samples from one distribution using another distribution. The KL divergence over the entire joint distribution is computationally prohibitive for most of the simulations covered in this section, therefore we compute the KL divergence over marginal distributions as follows:

$$\sum_{i=1}^{n}\sum_{t=0}^{T} \mathrm{KL}(P(X_i^{(t)}) \mid\mid Q(X_i^{(t)})) = \sum_{i=1}^{n}\sum_{t=0}^{T}\sum_{X_i} P(X_i^{(t)}) \log \frac{P(X_i^{(t)})}{Q(X_i^{(t)})} \tag{6.3}$$

where $Q(X_i^{(t)}) = 0$ implies $P(X_i^{(t)}) = 0$. Equation 6.3 corresponds to the sum of the divergences between the marginal distributions P and Q, in this case a reference distribution and a learned distribution respectively. As with the standard KL divergence, the quantity of Equation 6.3 should be minimized.

6.4.2 *Learning and evaluating PDBNs*

In order to learn a PDBN with k cuts, k homogeneous models are learned using the corresponding portions of the training data according to its cut set, where each sub-DBN is learned separately. As it happens with Bayesian-network

learning, typically search-and-score and constraint-based methods are used for learning the structure and parameters of each sub-DBN (see Section 2.4.2). In the experiments reported in this chapter, the AIC score (see Equation 5.2) is employed for evaluating each sub-DBN, which yields a score proportional to the likelihood of the model and a penalization term for the complexity.

In order to select a suitable number of cuts, we implemented the evaluation criterion of Algorithm 2 by means of a 10-fold cross-validation. Cross-validation minimizes the effect of overfitting (see Section 2.6.3); we describe the procedure in detail in the following. Let $C_i = \{t_1, \ldots, t_k\}$ be a cut set of a time series over $\{0, \ldots, T\}$; in the context of Algorithm 2, C_i corresponds to 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-k. After processing all the folds, the mean of the log-likelihoods is taken, which represents the evaluation value of the PDBN-k with cut set C_i, as indicated in Algorithm 2 by $f(C_i)$. When deciding between two cut sets (e.g. as in Line 5), the algorithm chooses the one having the higher mean log-likelihood.

After leaving the outer loop of Algorithm 2, the heuristic search is finished and the best cut set is known. Finally, a PDBN-k with such cut set is learned using the full dataset, i.e. training and test data. Such PDBN-k corresponds to the output of the procedure.

6.4.3 *Results and discussion*

The results of simulations with data generated from DBN, PDBN-2 and PDBN-3 models are shown in Tables 6.1, 6.2 and 6.3 respectively. Note that a DBN was learned on every case to serve as a baseline method, specially when simulating data from non-DBNs; the performance of the learned DBNs are indicated on the sixth column of the tables. Table 6.1 shows that the models learned by the heuristic based on DBN data have structural partitioning in accordance with the reference models on most cases, indicating that the heuristic was capable of retrieving the adequate type of model. When the returned models were not a DBN, they were mostly only 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 well as the learned DBNs did.

The models returned by the heuristic based on data produced by PDBNs-2 and PDBNs-3 (Tables 6.2 and 6.3) 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, DBNs capture the average behavior of 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 PDBN-2

n	d	Learned Model	Cut Sets (R; L)	Cut Diff.	KL(L DBN)	KL(L)
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

Table 6.1: Simulations with *data generated from DBNs*, where n and d denote the number of features and the number of sequences respectively. **R** = reference model, **L** = learned model (heuristic), **KL (M)** = KL divergence between model M and the reference model, **L DBN** = learned DBN.

and PDBN-3 cases where the first cut was situated around half of the sequence duration, there were at least two different transition patterns, which tends to make DBNs 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 just a few 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 made mistakes with low impact nonetheless.

n	d	Learned Model	Cut Sets (R; L)	Cut Diff.	KL(L DBN)	KL(L)
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 6.2: Simulations with *data generated from PDBN-2 models*. The best KL divergence values are given in bold face and followed by an asterisk.

A summary of the results presented in Tables 6.1, 6.2 and 6.3 is given in Table 6.4. Each row of the table aggregates simulations of DBNs, PDBNs-2 and PDBNs-3 according to the number of features and sequence length.

6.4.4 *Small datasets*

In the final analysis based on simulations, we focus on the small datasets. The simulations suggest that the models learned by the heuristic from the smallest

n	d	Learned Model	Cut Sets (R; L)	Cut Diff.	KL(L DBN)	KL(L)
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 6.3: Simulations with *data generated from PDBN-3 models*. The best KL divergence values are given in bold face and followed by an asterisk.

datasets (i.e. those with $d = 100$ sequences) were simpler than the reference models used to generate simulated data in virtually every case. Hence, the heuristic tends to operate in a conservative mode when there is scarcity of data. This also indicates that the methodology was effective in combating overfitting in 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 in 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 PDBN-3 models,

n	KL(DBN) - KL(L)	KL(L)	'$=$' (total)	'$\subseteq +a$'	'$\not\subseteq$' (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)

Table 6.4: Summary of simulations with DBNs and PDBNs. Abbreviations: **L** = learned model (heuristic), **KL (M)** = KL divergence between model M and the reference model. Positive values in the 2nd column indicate higher divergences achieved by DBNs. The 4th, 5th and 6th columns refer to the structural comparison of Section 6.4.1 and stand for the number of equal cut sets, average number of additional cut sets in learned models, and number of remaining cases respectively.

indicating a decent learning ability of the heuristic in the difficult situation of small datasets.

6.5 LEARNING TEMPORAL MODELS OF PSYCHOTIC DEPRESSION

6.5.1 *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 [41, 157] and eliciting both structure and parameters from descriptive statistics and expert knowledge [49, 103]. 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 [49, 103], 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 the literature on BNs in psychiatry, so far time has not been a factor that has been taken into account in a comprehensive manner. Except for [41] that deals with the beginning and end of treatment, research that considers a broad range of time granularities has not been done up to this moment. This could be of interest, e.g., to controlled treatment trials and longitudinal diagnosis, where the examination of some form of history or time series measurements would allow for a more global comprehension of, for example, the evolution of mental illnesses and a more accurate diagnosis. For prediction with BNs and extensions such as DBNs, it is not required to enter all the symptoms as input for these models to be able to deliver predictions about the future. Furthermore, these predictions can be done for any point in future. Besides prediction, temporal models can also be used to find associations taking into account the time dimension. On the other hand, well-known models such as regression seem to be less flexible with regard to tasks such as the mentioned ones.

Within the field of psychiatry, diseases that have been covered under a BN approach include depression [36, 41, 103], social anxiety [157], schizophrenia [49], as well as analyzing the use of BNs on diagnosis in psychiatry [162]. Moreover, there is little research on using temporal models for better understanding psychotic depression, which besides being a severe mental disorder, brings an additional complexity due to the presence of psychosis and depression factors.

6.5.2 *Problem description and data*

To illustrate the use of non-homogeneous probabilistic models and the heuristic construction procedure proposed in this work, a case study in psychiatry is considered. It comprises a dataset from an original study designed to assess three different drugs to treat psychotic depression over 7 weeks [175]. The primary outcome of the original study aimed at comparing the drugs to depression levels and psychotic features at treatment endpoint. In this work, we aimed 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 used to evaluate a large range of hypothesis about PD while modeling explicit relationships between psychotic and depressive features. We first discuss the results obtained by the heuristic algorithm when applied over psychiatry data, aiming at: (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 6.6 we make use of the obtained models to answer 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 (abbreviated as HDRS17) [81], in this section 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) [89] instead of the complete HDRS17, consisting of the features shown on Table 6.5. Using the melancholia sub-scale is, therefore, two-fold: it avoids

the usage of the complete HDRS17 upon the available scarce dataset, whereas HDRS6 is able to capture the core symptoms of depression [89]. In addition, two psychotic features were considered (hallucinations and delusions), totalizing eight features.

Psychiatry dataset [175]	
Number of sequences (complete)	122 (100) patients
Number of time points	8 (including baseline)
Depression features (HDRS6)	Depressed mood (Dm), Guilt (Gu), Work and Activities (Ac), Psychomotor Retardation (Re), Psychic Anxiety (Ap), and Somatic General (Sg)
Psychotic features	Hallucinations (Ha) and Delusions (De)
Study's period and location	2002-2007, The Netherlands

Table 6.5: Summary of psychiatry data.

The *somatic general* item takes values from the set $\{0, 1, 2\}$, 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* [81]. To use as much data as possible, the incomplete cases were imputed with the same method used in the original study [175], namely, the last observation carried forward (LOCF). The frequencies of the imputed data at each week are shown on Table 6.6. 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*.

6.5.3 *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 a better fit 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. 6.2. Although not expanded further, the model with cuts $\{6, 7\}$ was also a better fit 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. 6.2 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 best fitting PDBN-

	Depressed mood						Guilt					
t	**0**	**1**	**2**	**3**	**4**	μ	**0**	**1**	**2**	**3**	**4**	μ
0	0	0	0.04	0.35	0.61	**3.57**	0.04	0.05	0.14	0.14	0.63	**3.27**
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**
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**
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**
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**
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**
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**
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**

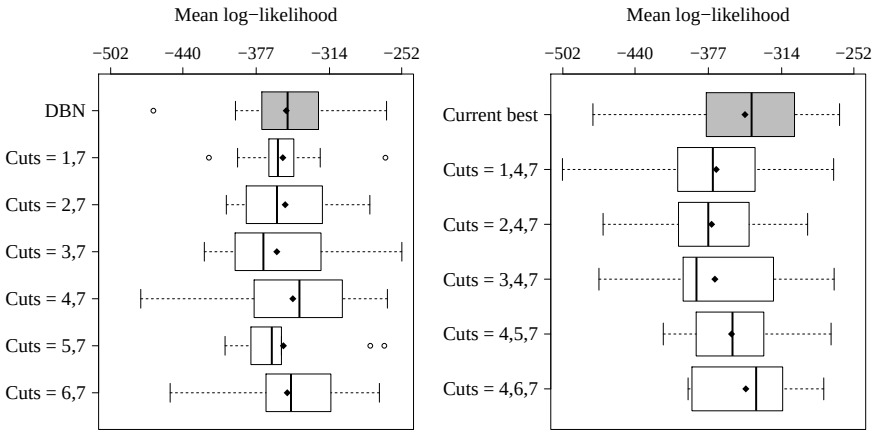
	Psychomotor retardation						Psychic anxiety					
t	**0**	**1**	**2**	**3**	**4**	μ	**0**	**1**	**2**	**3**	**4**	μ
0	0.16	0.3	0.31	0.22	0.02	**1.65**	0.03	0.14	0.27	0.37	0.19	**2.54**
1	0.15	0.33	0.34	0.16	0.02	**1.59**	0.11	0.16	0.29	0.29	0.16	**2.22**
2	0.27	0.3	0.29	0.12	0.02	**1.34**	0.18	0.22	0.3	0.23	0.07	**1.8**
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**
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**
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**
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**
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**

	Work and activities						Somatic general			
t	**0**	**1**	**2**	**3**	**4**	μ	**0**	**1**	**2**	μ
0	0	0	0.15	0.49	0.36	**3.21**	0.1	0.3	0.61	**2.54**
1	0	0	0.21	0.52	0.27	**3.06**	0.16	0.34	0.51	**2.22**
2	0	0.02	0.34	0.5	0.14	**2.76**	0.22	0.43	0.34	**1.8**
3	0.01	0.08	0.35	0.4	0.16	**2.61**	0.34	0.39	0.27	**1.47**
4	0.02	0.12	0.4	0.34	0.12	**2.43**	0.27	0.48	0.25	**1.42**
5	0.02	0.14	0.43	0.29	0.12	**2.34**	0.39	0.42	0.2	**1.24**
6	0.03	0.19	0.36	0.3	0.12	**2.29**	0.41	0.38	0.21	**1.3**
7	0.07	0.25	0.37	0.2	0.11	**2.03**	0.4	0.36	0.24	**1.2**

Table 6.6: Relative frequencies of HDRS6 items of psychiatry data at each week, where μ denotes the respective weighted means.

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.



(a) First iteration: 2 cuts.

(b) Second iteration: 3 cuts.

Figure 6.2: Boxplots for each stage of the heuristic over psychiatry data. The means are represented by a diamond symbol.

6.5.4 *Transition structures*

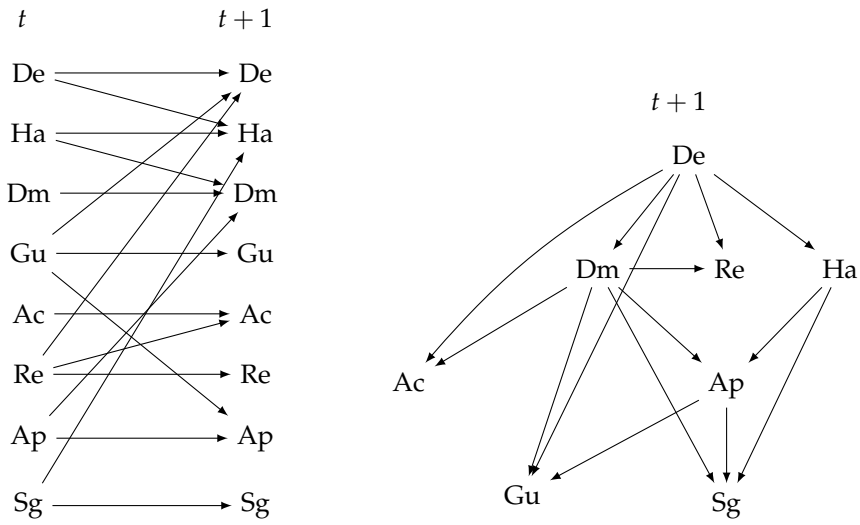
The structure of the DBN is shown in Fig. 6.3, while the structure of the conditional BNs that compose the PDBN-3 are shown in Figures 6.4 and 6.5. For a clearer exposition, each conditional BN was split into *inter*-temporal arcs (i.e. those from $t+1$ to t) and *intra*-temporal arcs (those delimited 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 $A^{(t+1)}$.

6.6 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 clinically-oriented questions.

6.6.1 *Marginals of symptoms 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 6.7 presents the empirical and model-based marginals for each symptom



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

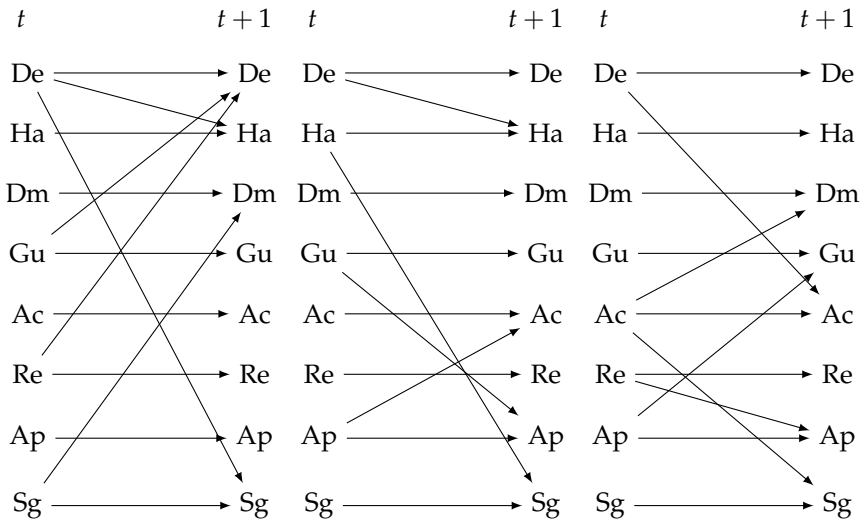
Figure 6.3: Structure of the DBN learned from the psychiatry data. Nodes on the left side of the inter-temporal arcs occur at time t, while those on the right at $t+1$. De = Delusions, Ha = Hallucinations, Dm = Depressed mood, Gu = Guilt, Ac = Work and activities, Re = Psychomotor retardation, Ap = Psychic anxiety, Sg = Somatic general.

per week, where the value assumed is either *true* or *high*. A summary of this information is presented at Table 6.8.

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.

6.6.2 *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 [104], essentially testing the paths between the respective nodes in the structure. As the Figures 6.4 and 6.5 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.

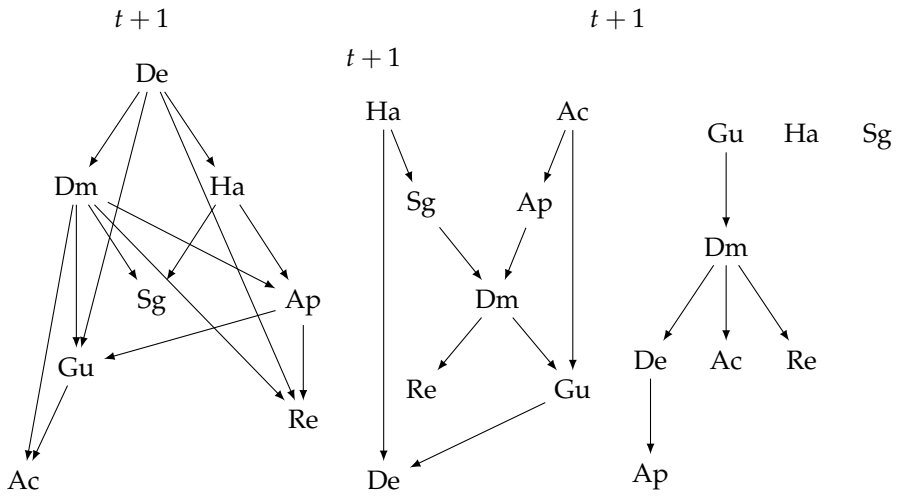


(a) First cut: $[0, 4]$. (b) Second cut: $[4, 6]$. (c) Third cut: $[6, 7]$.

Figure 6.4: *Inter*-temporal arcs of the PDBN-3 learned from the psychiatry data. De = Delusions, Ha = Hallucinations, Dm = Depressed mood, Gu = Guilt, Ac = Work and activities, Re = Psychomotor retardation, Ap = Psychic anxiety, Sg = Somatic general.

In this context, we approach a research question within the field of psychiatry, specially in psychotic depression: *to which extent do psychotic and depressive features 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 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.



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

Figure 6.5: *Intra*-temporal arcs of the PDBN-3 learned from the psychiatry data. De = Delusions, Ha = Hallucinations, Dm = Depressed mood, Gu = Guilt, Ac = Work and activities, Re = Psychomotor retardation, Ap = Psychic anxiety, Sg = Somatic general.

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, $\text{dom}(X) = \{true, false\}$ and $\text{dom}(Y) = \{low, high\}$. Then, the odds ratio to predict Y given X is:

$$\text{OR}(Y^{(i)}|X^{(1)}) = \frac{\text{odds}(Y^{(i)} = high \mid X^{(1)} = true)}{\text{odds}(Y^{(i)} = high \mid X^{(1)} = false)} \tag{6.4}$$

$$= \frac{\dfrac{P(Y^{(i)} = high \mid X^{(1)} = true)}{1 - P(Y^{(i)} = high \mid X^{(1)} = true)}}{\dfrac{P(Y^{(i)} = high \mid X^{(1)} = false)}{1 - P(Y^{(i)} = high \mid X^{(1)} = false)}} \tag{6.5}$$

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 feature X. If OR > 1, then it is more likely that the depressive feature 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 OR < 1, it is more likely to observe Y at *high* in the group $X = false$ than in the group $X = true$; finally, if OR $= 1$, there is no association between X and Y, i.e. knowing the group of this particular psychotic feature does not affect the predictions for this depressive symptom. For the sake of terminology, an OR > 1 is also called a positive correlation, while an OR < 1 indicates a negative correlation. Note that

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 6.7: 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.

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 $\alpha = 0.05$.

6.6.2.1 *Predictors for depression*

Table 6.9 shows the ORs for psychotic features 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

Symptom	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***

Table 6.8: Summary of percentage differences of learned models to the marginal frequencies of psychiatry data. The absolute values are used to compute the means.

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).

With respect to the predictive power of psychotic symptoms observed at $t = 4$ and $t = 6$ (Table 6.10, 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.

Symptom & predictor	t=2	t=3	t=4	t=5	t=6	t=7
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

Table 6.9: Odds ratios for **psychotic symptoms as predictors**. An OR greater than 1 indicates that the level *high* on the depressive feature is more likely to be observed in the group *true* than in the group *false* of the psychotic feature. Results marked in bold and * stand for a statistically significant association.

Symptom & predictor	t=5	t=6	t=7
Depressed mood			
Delusions$^{(4)}$	**3.09***	**2.26***	**2.17***
Hallucinations$^{(4)}$	1.98	2.14	1.71
Guilt			
Delusions$^{(4)}$	**4.15***	**2.93***	**2.34***
Hallucinations$^{(4)}$	1.19	1.31	1.4
Activities			
Delusions$^{(4)}$	2.26	1.81	**2.59***
Hallucinations$^{(4)}$	2.52	1.53	1.61
Retardation			
Delusions$^{(4)}$	**2.97***	2.02	1.98
Hallucinations$^{(4)}$	1.4	1.25	1.36
Psychic anxiety			
Delusions$^{(4)}$	1.88	1.88	**2.21***
Hallucinations$^{(4)}$	2.18	1.53	1.45
Somatic general			
Delusions$^{(4)}$	0.97	0.87	0.99
Hallucinations$^{(4)}$	**6.52***	**6.18***	**4.91***

Symptom & predictor	t=7
Depressed mood	
Delusions$^{(6)}$	**2.72***
Hallucinations$^{(6)}$	1.67
Guilt	
Delusions$^{(6)}$	**3.62***
Hallucinations$^{(6)}$	1.2
Activities	
Delusions$^{(6)}$	**5.66***
Hallucinations$^{(6)}$	1.61
Retardation	
Delusions$^{(6)}$	2.04
Hallucinations$^{(6)}$	1.34
Psychic anxiety	
Delusions$^{(6)}$	**3.52***
Hallucinations$^{(6)}$	1.25
Somatic general	
Delusions$^{(6)}$	1.14
Hallucinations$^{(6)}$	**4.31***

Table 6.10: Odds ratios for **psychotic symptoms as predictors** (cont.). Left: $t = 4$, right: $t = 6$.

6.6.2.2 *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 6.11a 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 $t = 1$.

On the other hand, updating the depressive symptoms at $t = 4$, as shown on Table 6.11b (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.

6.7 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

Symptom & predictor	t=2	t=3	t=4	t=5	t=6	t=7
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

(a) Odds ratios based on $t = 1$.

Symptom & predictor	t=5	t=6	t=7
Delusions			
Depressed mood$^{(4)}$	**4.96***	**3.97***	**4.22***
Guilt$^{(4)}$	**8.13***	**5.62***	**4.58***
Activities$^{(4)}$	3.84	3.36	3.14
Retardation$^{(4)}$	**3.32***	**2.5***	**2.2***
Psychic anxiety$^{(4)}$	1.8	1.69	1.9
Somatic general$^{(4)}$	1.35	1.35	1.37
Hallucinations			
Depressed mood$^{(4)}$	1.2	1.2	0.97
Guilt$^{(4)}$	1.07	0.91	0.96
Activities$^{(4)}$	0.82	0.9	0.67
Retardation$^{(4)}$	1.04	1.3	1.27
Psychic anxiety$^{(4)}$	**3.8***	**3***	2.97
Somatic general$^{(4)}$	**4.85***	**3.6***	**3.36***

Symptom & predictor	t=7
Delusions	
Depressed mood$^{(6)}$	**3.94***
Guilt$^{(6)}$	**5.63***
Activities$^{(6)}$	1.83
Retardation$^{(6)}$	1.87
Psychic anxiety$^{(6)}$	**2.52***
Somatic general$^{(6)}$	1.19
Hallucinations	
Depressed mood$^{(6)}$	1.71
Guilt$^{(6)}$	1.35
Activities$^{(6)}$	1.25
Retardation$^{(6)}$	1.41
Psychic anxiety$^{(6)}$	1.29
Somatic general$^{(6)}$	**7.47***

(b) Odds ratios based on $t = 4$ (left) and $t = 6$ (right).

Table 6.11: Odds ratios for **depressive symptoms as predictors**. An OR greater than 1 indicates that the level *true* on the psychotic feature is more likely to be observed in the group *high* than in the group *low* of the depressive feature. Results marked in bold and * stand for a statistically significant association.

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, simulated 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 aspects (e.g. model fit and problem insight) when the underlying process switches between different regimes on time. In the case of small datasets, common in many clinical studies, it was shown 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 employed by the heuristic explicitly

avoids over-specialized models, at the same time providing more robust models. Moreover, the search strategy of the heuristic, based on incremental construction of non-homogeneous models, is able to 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 patient recovery in psychotic depression over the course of a controlled treatment. In particular, a research question aiming to answer the temporal relationship between psychotic and depressive features 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 is 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 in 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 by choosing one of these new solutions randomly and then resuming the search. The problem of handling multiple solutions is in fact recurring in the literature of Bayesian networks, where extensive research has been developed [33, 40, 108, 124]. In this direction, the approach of this chapter could benefit from such research, for example by extending the greedy search, as well as taking into account Bayesian approaches [145]. These further investigations could provide more insight about the distribution and the variance of the cut sets.