Modeling Dynamic Systems with Memory: What Is the Right Time-Order?

Anna Łupińska-Dubicka[1] and **Marek J. Druzdzel**[1,2]
a.lupinska@pb.edu.pl, marek@sis.pitt.edu
[1] Faculty of Computer Science, Białystok University of Technology, Wiejska 45A, 15-351 Białystok, Poland
[2] Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program,
University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract

Most practical uses of Dynamic Bayesian Networks (DBNs) involve temporal influences of the first order, i.e., influences between neighboring time steps. This choice is a convenient approximation influenced by the existence of efficient algorithms for first order models and limitations of available tools. We focus on the question whether constructing higher time-order models is worth the effort when the underlying system's memory goes beyond the current state. We present the results of an experiment with a series of DBN models monitoring woman's monthly cycle. We show that higher order models are significantly more accurate. However, we have also observed overfitting and a resulting decrease in accuracy when the time order chosen is too high.

1 Introduction

All real world systems change over time. Modeling their equilibrium states or ignoring change altogether, when it is sufficiently slow, is sufficient for solving a wide spectrum of practical problems. In some cases, however, it is necessary to follow the change that the system is undergoing and introduce time as one of the model variables.

We concentrate in this paper on models that belong to the class of probabilistic graphical models, with their two prominent members, Bayesian networks (BNs) (Pearl, 1988) and dynamic Bayesian networks (DBNs) (Dean & Kanazawa, 1989). BNs are widely used practical tools for knowledge representation and reasoning under uncertainty in equilibrium systems. DBNs extend them to time-dependent domains by introducing an explicit notion of time and influences that span over time. Most practical uses of DBNs involve temporal influences of the first order, i.e., influences between neighboring time steps. This choice is a convenient approximation influenced by existence of efficient algorithms for first order models and limitations of available tools. After all, introducing higher order temporal influences may be costly in terms of the resulting computational complexity of inference, which is NP-hard even for static models. Limiting temporal influences to influences between neighboring states is equivalent to assuming that the only thing that matters in the future trajectory of the system is its current state. Many real world systems, however, have memory that spans beyond their current state.

The question that we pose in the paper is whether introducing higher order influences, i.e., influences that span over multiple steps, is worth the effort in the sense of improving the accuracy of the model. The idea of increasing modeling accuracy by means of increasing the time order of the model was beautifully illustrated by Shannon (1948). In his seminal paper, he shows sentences in the English language, generated by a series of Markov chain models of increasing time order, trained by means of the same corpus of text. The following sentence was generated by a first order model:

```
OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH
EEI ALHENHTTPA OOBTTVA NAH BRL.
```

Compare this with the following sentence generated by a sixth order model:

```
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE LETTERS
THAT THE TIME OF WHO EVER TOLD THE PROBLEM
FOR AN UNEXPECTED.
```

The resemblance of the latter sentence to ordinary English text, an informal measure of the model's accuracy, has increased dramatically between the first and the sixth orders. A first order model was essentially impotent in its ability to model the problem.

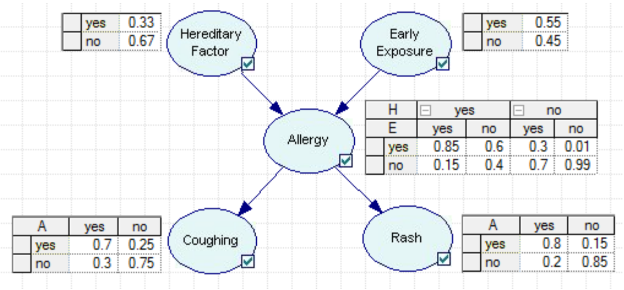
While generation of English sentences may be too hard of a problem, the vehicle for our experiments is the problem of monitoring the woman's monthly cycle, a problem central to family planning. Every couple seeking help in a fertility clinic is asked to monitor the monthly cycle before any intervention is undertaken. An accurate monitoring model can be a great aid in natural family planning, indicating optimal days for sexual intercourse. What is important from the perspective of the question posed in this paper is that woman's monthly cycle is a system with memory going most certainly beyond one day and probably spanning over a period of roughly a month.

We report the results of an experiment in which we successively introduce higher order DBNs modeling the monthly cycle and measure the accuracy of these models in predicting the day of ovulation. We train our models on real time series data obtained from a longitudinal study of fecundability conducted in several European centers (Colombo & Masarotto, 2000). We show that increasing the time order of the model greatly improves its accuracy. However, we also observe that when the time order is too high, the model can overfit the data and the quality of its predictions may decrease.

2 BNs and DBNs

Consider the simple BN shown in Figure 1, illustrating various causes and effects of allergy in children. All variables in this example are Boolean. The tendency to develop allergies has a hereditary component: Allergic parents are more likely to have allergic children, whose allergies are likely to be more severe than those from non-allergic parents. Exposure to allergens, especially in early life, is also an important risk factor for allergy. When an allergen enters the body of an allergic child, the child can cough or develop a rash. Figure 1 shows the dependency structure among the variables and the conditional probability distributions for each of the variables.

DBNs (Dean & Kanazawa, 1989) are an extension of BNs for modeling dynamic systems. The term *dynamic* means that we model the system's development over time and not that the model structure and its parameters change over time, even though the latter is theoretically possible. In a DBN, the state of a system at time t is represented by a set of random variables $\mathbf{X}^t = (X_1^t, \ldots, X_n^t)$. The state at time t generally dependents on the states at previous k time steps. There is nothing in the theory that prevents k from being any number between 1 and $t-1$.



yes	0.33
no	0.67

yes	0.55
no	0.45

H	yes		no	
E	yes	no	yes	no
yes	0.85	0.6	0.3	0.01
no	0.15	0.4	0.7	0.99

A	yes	no
yes	0.7	0.25
no	0.3	0.75

A	yes	no
yes	0.8	0.15
no	0.2	0.85

Figure 1: A simple BN illustrating selected causes and effects of allergy in children

When each state of the model depends only on the immediately preceding state (i.e., $k = 1$, the system is first-order Markov, often assumed in practice), we represent the transition distribution $P(X^t | X^{t-1})$. This can be done using a two-slice BN fragment (2TBN) $\mathcal{B}^t$, which contains variables from $\mathbf{X}^t$ whose parents are variables from $\mathbf{X}^{t-1}$ and/or $\mathbf{X}^t$, and variables from $\mathbf{X}^{t-1}$ without their parents. A first order DBN is often defined as a pair of BNs $(\mathcal{B}^0, \mathcal{B}^{\rightarrow})$, where $\mathcal{B}^0$ represents the initial distribution $P(\mathbf{X}^0)$, and $\mathcal{B}^{\rightarrow}$ is a two time slice BN, that defines the transition distribution $P(\mathbf{X}^t | \mathbf{X}^{t-1})$ as follows:

$$P(\mathbf{X}^t | \mathbf{X}^{t-1}) = \prod_{i=1}^{n} P(X_i^t | Pa(X_i^t)) \ .$$

Consider a two years old child whose parents suffer from allergy and who has been exposed to allergens. We know that this child has not developed any symptoms of allergy in the previous year. Suppose that we want to know the probability that allergy appears in the third year. If we use the BN pictured in Figure 1, we omit all historical information except that for the current year. Figure 2a shows a DBN of first temporal order, which allows us to predict the probability of the child developing allergy in this and in the future years. Number of slices is the number of steps for which we perform the inference. In this example, one step means one year. Temporal plate is the part of a DBN that contains nodes changing over time. *Hereditary Factor* is outside of the temporal plate and, hence, is time invariant.

Figure 2b shows a second time-order DBN, i.e., a model in which there are two temporal arcs from node *Allergy*, the first order takes the information from one step before, the second from two steps before. Typically, the older the child, the lower the probability of allergy appearing. And, generally, a child who has not developed allergy two years in a row has a lower chance of developing allergy in the third year. A reasonable expectation is that modeling higher order dependencies should increase the accuracy of the model.

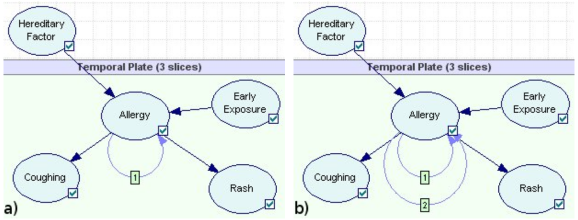


Figure 2: DBNs modeling causes and effects an allergy in children: first order (a) and second order (b) DBN

3 Woman's monthly cycle

Woman's monthly cycle is driven by a highly complex interaction among hormones produced by three organs of the body: the hypothalamus, the pituitary gland, and the ovaries. There are five main hormones involved in the menstrual cycle process: estrogen, progesterone, gonadotropin releasing hormone (GnRH), follicle stimulating hormone (FSH), and lutenizing hormone (LH).

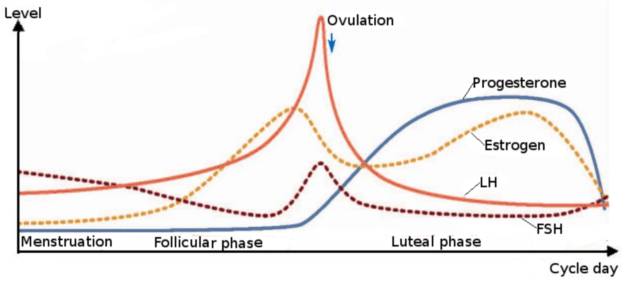


Figure 3: Levels of hormones during the phases of the woman's monthly cycle (Barron & Fehring, 2005)

The woman's monthly cycle consists of four phases (Figure 3 shows these four phases along with the associated hormone levels): (1) menstruation, (2) the follicular phase, (3) ovulation, and (4) the luteal phase. Counting from the first day of the menstrual flow, the length of each phase may vary from woman to woman and then cycle to cycle.

In addition to measurable blood hormone levels, there are several easily accessible indicators of the phase of the cycle, two of which we will use in our models. Basal body temperature (BBT) is defined as the body temperature measured immediately after awakening and before any physical activity has been undertaken. It should be measured every day at the same time. Before ovulation, BBT is relatively low. Following the ovulation, as a result of an increased level of progesterone in the body, women typically experience an increase in the basal body temperature (BBT) of at least 0.2°C. This shift indicates that ovulation has occurred. The BBT charting may provide valuable information about woman's monthly cycle, such as duration of the cycle, length of the follicular and luteal phases, and the pattern of the timing of ovulation. Sometimes BBT can rise due to causes other than ovulation. This atypical rise is treated as disturbance and can be caused by a change in conditions around the measurement, such as later measurement time, lack of sleep, different thermometer, high stress, travel, or illness.

As the cycle progresses, due to hormonal fluctuations, the cervical mucus increases in volume and changes texture. When there is no mucus or the mucus discharge is small, the day is considered infertile. There can be also a feeling of dryness around the vulva. Around the ovulation, mucus is the thinnest, clearest, and most abundant, resembling egg white. In the luteal phase, it returns to the sticky stage.

It seems that the menstrual cycle is a temporal process with memory spanning over the entire cycle. This means that the current state is not only influenced by the previous state but also by prior days, going back to the beginning of the phase.

4 The Model

Accurate prediction of the fertile phase of the menstrual cycle is critical for couples who want to conceive or couples who want to avoid pregnancy using natural methods. The fertile phase of the menstrual cycle is defined as the time when an intercourse has a non-zero probability of resulting in conception.

The number of fertile days during the menstrual cycle is difficult to specify, as it depends on the life span of the ovum and sperm, which varies from person to person and from cycle to cycle. It is generally believed that an ovum can be fertilized only within the first 24 hours after ovulation. Many authors agree that the start of the fertile interval is strictly connected with changes in vaginal discharge and, in particular, estrogenic-type cervical mucus secretions. However, they differ in their estimates of the length of the fertile window. Potter (1961) calculated that there are only two days during the menstrual cycle when a woman can become pregnant. Wilcox *et al.* (1995) found that the maximum sperm life span equals approximately five days (in presence of sufficient level of estrogenic-type mucus), which comes down to a fertile period of six days, including the day of the ovulation. The results of a multi-center study conducted by the World Health Organization (WHO, 1983) estimate the fertile period to be 10-days before ovulation. Natural family planning methods assume this interval to be as long as 13 days.

It is useful and important to be able to predict ovulation. Because the fertile period starts roughly five days before ovulation, prediction has to be made in advance and, hence, asks for models that include an explicit notion of time.

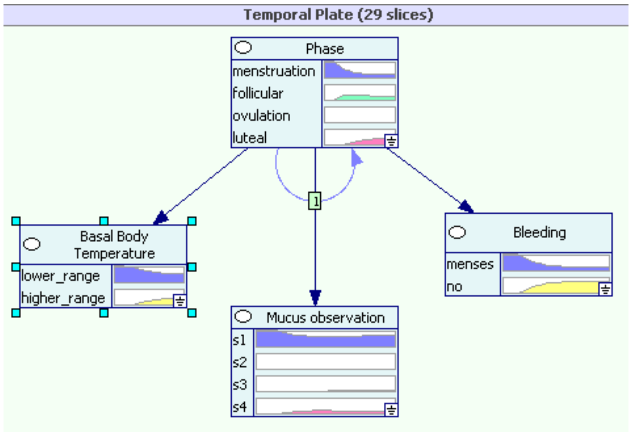


Figure 4: A first-order DBN model of woman's monthly cycle

Our model (Figure 4), combines information retrieved from BBT charting with observations of the cervical mucus secretions. It contains a variable *Phase* with four states: menstruation, follicular, ovulation, and luteal. We included three observation variables: *Basal Body Temperature* (BBT), *Bleeding* and *Mucus observation*. All variables are discrete. BBT has two possible values: lower range and higher range, representing temperature before and after the BBT shift respectively. *Bleeding* describes whether on a particular day the woman had menses or not. *Cervical observation* can be in one of four states (s1 through s4), described in detail in (Dunson, Sinai, & Colombo, 2001). We modeled time explicitly as n time steps, where n is the number of days of the longest monthly cycle of the particular woman.

Admittedly, this is a simple model. However, we would like to point out that it reasonably models the causal interactions among the variables in the data available to us.

5 The Training Data

Our training data are drawn from an Italian study of daily fecundability (Colombo & Masarotto, 2000), which enrolled women from seven European centers (Milan, Verona, Lugano, Düsseldorf, Paris, London and Brussels). To our knowledge, this is one of the most comprehensive data sets describing woman's monthly cycle. Between the years 1992 and 1996, 782 women recorded a total of over six thousand monthly cycles. Women participating in the study satisfied the following five entry criteria: (1) experienced in use of a Natural Family Planning method, (2) married or in a stable relationship, (3) between 18th and 40th birthday at admission, (4) had at least one menses after cessation of breastfeeding or after delivery, (5) not taking hormonal medication or drugs affecting fertility. In addition, neither partner could be permanently infertile and both had to be free from any illness that may affect fertility.

In each menstrual cycle, the subject was asked to record the days of her period, her basal body temperature and any disturbances such as illness, disruption of sleep or travel. She was also asked to observe and chart her cervical mucus symptoms daily during the cycle and to record every episode of coitus, with specification whether the couple used contraceptives or not.

Typically, a menstrual cycle is defined as the interval in days between the first day of menstrual bleeding in two neighboring cycles, where day 1 was the first day of fresh red bleeding, excluding any preceding days with spotting. The day of ovulation was identified in each cycle from records of basal body temperature and mucus symptoms. The daily mucus observations were classified into four classes; ranging from a score of 1 (no discharge and dry) to 4 (transparent, stretchy, slippery). The cervical mucus peak day was defined as the last day with best quality mucus, in a specific cycle of the woman. If there were different mucus observations on one day, the most fertile characteristic of the mucus observed determined the classification. To determine the BBT shift, the "three over six" rule was used: The first time in the menstrual cycle when three consecutive temperatures were registered, all of which were above the average temperature of the last six proceeding days.

6 Experiments

We tested our model on two different women taken from the Italian study. For each woman, we created a BN and nine DBNs of temporal orders ranging from 1 to 9, training them (i.e., learning their parameters) on the available monthly charts, using the leave-one-out method, i.e., training the network on all but one chart and testing it on the remaining chart. Because of computational limitations (with 30 time slices, ninth order models become fairly complex), we had to find women with a not too long average duration of the follicular phase. We were able to find two women with over 30 monthly charts each, whose follicular phase lasted typically around 9 days. We set the number of slices of the DBNs to the length of the longest cycle.

Just to give an idea of the capability of such models to reproduce the monthly cycle, we present the prob-

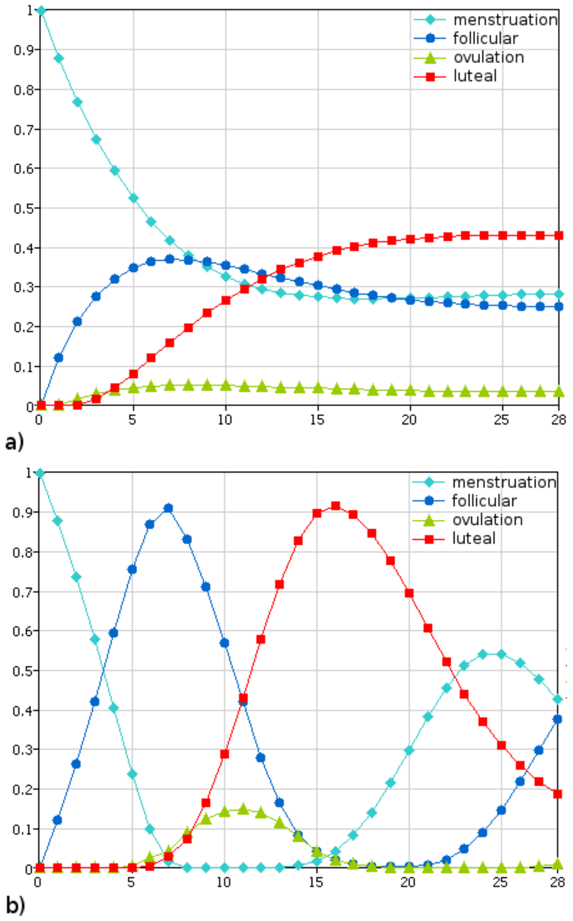


Figure 5: Probabilities of each phase during the monthly cycle: order 1 (a) and order 7 (b) DBNs

abilities of the four phases of the monthly cycle as a function of time in Figure 5. These probabilities were generated by models of the first (a) and the seventh (b) order DBNs, trained on monthly charts of one of the women in the data set. We entered no observation into the models, except for anchoring the first time step to the first day of menses, i.e., first day of the monthly cycle. Please note the increased similarity of the shape of the curves to that of the hormone levels in Figure 3, which are direct indications of phases of the monthly cycle. Memory of the order 7 model is such that the model is capable of predicting roughly the day of ovulation on the first day of menses.

To compare the accuracy of different models, we used two measures: the true positive rate (TPR) and the false positive rate (FPR). These are defined as $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$ respectively. In our model, TP is the number of true positives, i.e., the number of days of the cycle classified as ovulation that in fact were ovulation. FP is the number of false positives, i.e., the number of days of the cycle classified as ovulation that in fact belong to menstruation, follicular, or lutheal phase. TN is the number of true negatives, i.e., the number of days of the cycle not classified as ovulation that in fact belong to menstruation, follicular, or lutheal phase. FN is the number of false negatives, i.e., the number of days of the cycle not classified as ovulation that are ovulation.

From the practical perspective, for a model of a monthly cycle to be useful, it has to predict the day of ovulation at least five days in advance. Please note that because of a possible application of a model like this in family planning, false negatives may be very costly, so the model should minimize its false negative rate to zero. This is essentially the case with all natural family planning methods.

For each network, we created ROC graphs (Fawcett, 2003) by plotting sensitivity (TPR) vs. complement of specificity ($1-\mathrm{FPR}$). For each model, we had as many curves as there were cycles of the particular woman available. To plot the ROC curves, we used vertical averaging, i.e., for each FPR we took the averaged TPRs of the ROC curves over all cycles. For each curve, we also calculated the area under the ROC curve (AUC), which is a measure of model's ability to predict ovulation five days in advance. A useless model would have the AUC of 0.5. A model with perfect ability to predict would have the AUC of 1.0. If the 95% confidence interval of the model's AUC would include 0.5, the model would be not likely to predict accurately. We used ROCR (Sing, Sander, Beerenwinkel, & Lengauer, 1975), an R package for evaluating and visualizing classifier performance.

Figure 6 shows selected ROC curves created for the two selected women: static BN, first order DBN, DBNs with temporal orders from first to fourth, from first to sixth, from first to seventh, and with temporal orders from first to ninth. We did not picture every curve in order to avoid clattering the graphs but instead showed the ranges (vertical lines on the plot). Figure 7 presents the average AUCs for these women along with their ranges (vertical bars).

As we can see, in both women, a BN is not much better than a random classifier. From all DBNs, the networks with first temporal order and with first and second temporal orders give the worst results. In case of woman ID 20050265, the higher order of the network, the higher sensitivity at the same point of specificity (Figure 6a).

Figure 6b shows that for woman ID 20380003 the curve for DBN with orders higher than 6 does not achieve value $\mathrm{TPR} = 1.0$ until $1-\mathrm{FPR} = 0.42$. Starting at the $1-\mathrm{FPR} = 0.28$, these high order DBNs give worse results than DBNs with lower temporal orders. Fig-

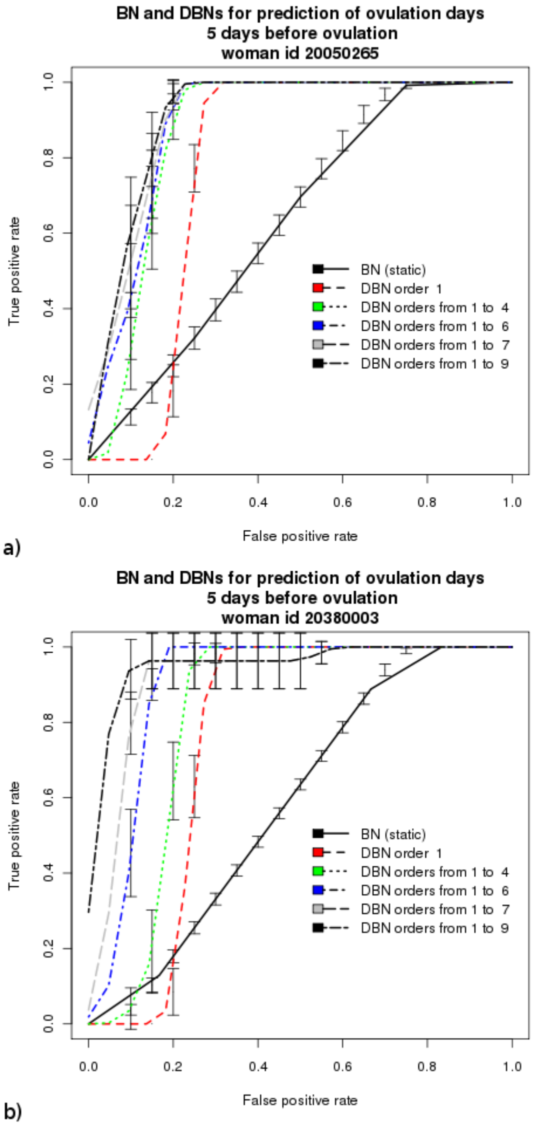


Figure 6: ROC curves with vertical averaging of BN and DBNs for prediction of the ovulation day

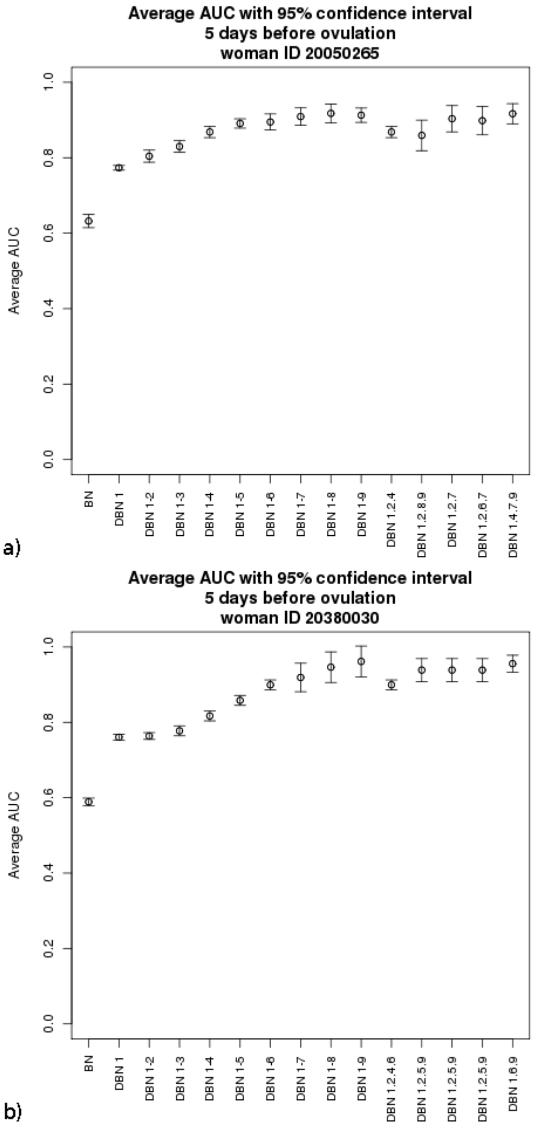


Figure 7: AUC ROC curves of BN and DBNs for prediction of the ovulation day

ure 8 shows this for each cycle separately. As we can see, there is one curve, whose AUC is smaller than 0.5. In this cycle, the follicular phase lasted only six days, while in all previous cycles its most common length was nine days. The model, learned on the basis of previous cycles, predicted ovulation day for the 15th day, while in reality it took place on the 12th day. Figure 9 is an equivalent of Figure 6b but with this anomalous cycle omitted. In this case, the higher order of the network, the higher sensitivity at the same point of specificity.

Clearly, too high of an order can reduce accuracy of the model. What is the optimal order of a model? We performed a number of additional experiments with monthly cycles of other women, varying the model order, and came to the conclusion that the optimal order of the DBN model depends directly on the nature of the system and the task that we set to perform. This number should be derived from the domain knowledge. If anomalies are to be expected, it does not make any sense to go beyond the order equal to the smallest of the following three numbers: (1) the length of the system's memory, which could be argued in our case to be the length of the woman's monthly cycle, (2) the length of the prediction horizon, which is the number of slices that we want to predict ahead (6 in our case), and (3) the maximum order that is still computable comfortably, which was in case of SMILE around 9.

Furthermore, while any DBN model should contain at least one first order influence (if that were not the case,

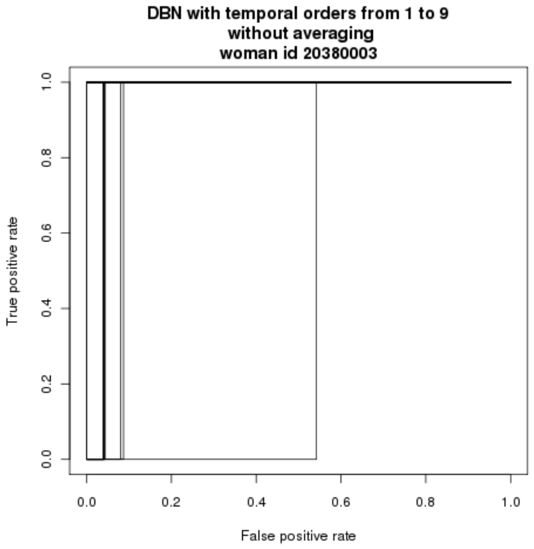


Figure 8: ROC curves for individual cycles of a DBN of order 1 through 9

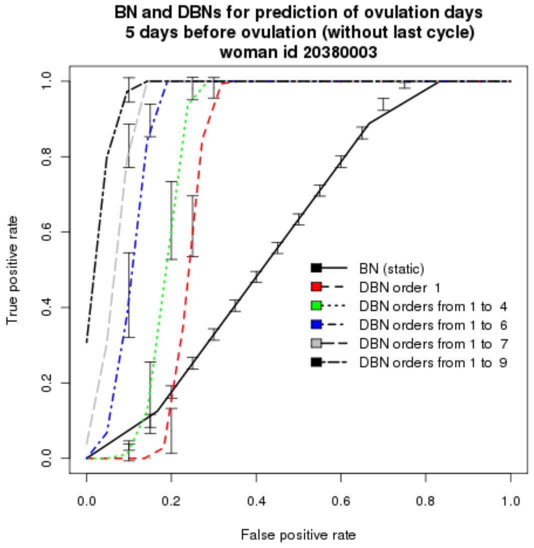


Figure 9: ROC curve for the DBN with temporal orders from 1 to 9 with anomalous cycle removed

some slices would be disconnected from the model!), a model of order k does not need to include influences of all orders between 1 and $k-1$. In our experiments with the monthly cycle, we focused on those influences that seemed critical for phase transitions and used orders that were equal to the lengths of the menstruation and the follicular phases, as given usually a clear indication the end of the menses, these influences could fairly precisely pinpoint the expected day of ovulation, even without additional observations.

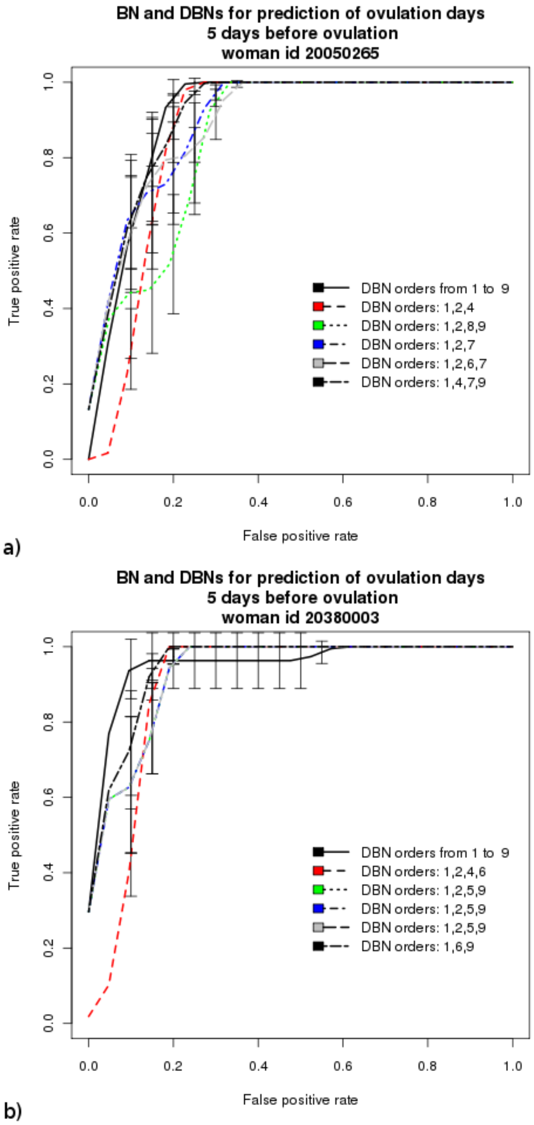


Figure 10: ROC curves with vertical averaging of additional DBNs for prediction of the ovulation day

Figure 10 shows ROC curves generated by DBNs with temporal orders including the shortest, the longest, the most common, and the average length of the menstruation and the follicular phases. The last pictured networks have temporal orders connected with the minimum, maximum, mode, and average length of the follicular phase of the particular woman, whose charts were used to train the model. Figure 11 shows networks with selected orders for woman 20380003 with the anomalous cycle removed.

7 Discussion

We have presented the results of an experiment with a series of DBN models monitoring woman's monthly cycle. We have shown that higher order models are

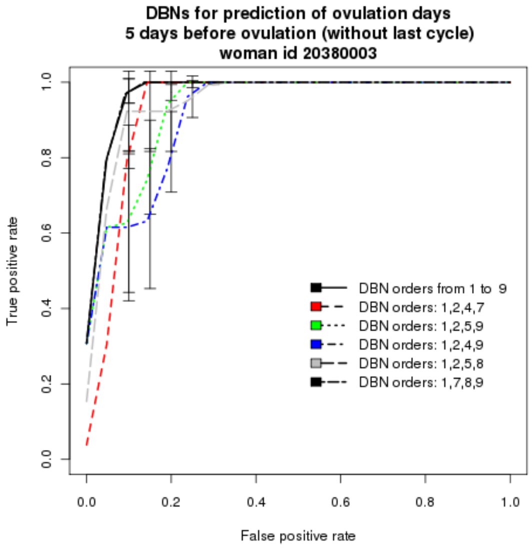


Figure 11: ROC curves with vertical averaging of additional DBNs for prediction of the ovulation day with the anomalous cycle removed

significantly more accurate than first order models, as summarized by the AUC graph in Figure 7. The ROC curves for higher order models were clearly closer to the upper left corner of an ROC graph, which indicates a better ability of the model to predict ovulation.

However, we also observed overfitting and a resulting decrease in accuracy when the time order chosen was too high. Having learned the lengths of the phases, which were shorter than the model's memory, the model seemed to lose its ability to predict accurately, when the cycle happened to be anomalous in terms of its length. Model's memory seemed to have a stronger influence on prediction than observations collected during the cycle. DBNs of lower orders reached sensitivity of 1.0 for lower values of specificity (Figures 6b and 10).

Thorough understanding of the underlying system and the task at hand is required to select the optimal order of the model. In addition to computational issues and issues related to a negative influence of model complexity on the quality of parameters learned from data, one should avoid choosing orders that are higher than system's memory and the task horizon.

Acknowledgements

Our work was supported in part by the National Institute of Health under grant number U01HL101066-01. We thank Bernardo Colombo, Guido Masarotto, Fausta Ongaro, Petra Frank-Herrmann, and other investigators of the European Study of Daily Fecundability for sharing their data with us. The empirical part of the paper was performed using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory, University of Pittsburgh, and available at `http://genie.sis.pitt.edu/`.

References

Barron, M. L., & Fehring, R. J. (2005). Basal body temperature assessment: Is it useful to couples seeking pregnancy? *American Journal of Maternal Child Nursing*, *30*(5), 290–296.

Colombo, B., & Masarotto, G. (2000). Daily fecundability: First results from a new data base. *Demographic Research*, *3*(5).

Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, *5*(2), 142–150.

Dunson, D. B., Sinai, I., & Colombo, B. (2001). The relationship between cervical secretions and the daily probabilities of pregnancy effectiveness of the TwoDay Algorithm. *Human Reproduction*, *16*(11), 2278–2282.

Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for researchers* (Technical Report No. HPL-2003-4). Hewlett Packard Laboratories.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Potter, J., R.G. (1961). Length of the fertile period. *Milbank Quarterly*, *39*, 132–162.

Shannon, C. E. (1948, July, October). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (1975). ROCR: Visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, *55*(4), 699–706.

WHO. (1983). A prospective multicentre trial of the ovulation method of natural family planning. III. Characteristics of the menstrual cycle and of the fertile phase. *Fertility and Sterility*, *40*(6), 773–778.

Wilcox, A., Weinberg, C., & Baird, D. (1995). Timing of sexual intercourse in relation to ovulation. Effects on the probability of conception, survival of the pregnancy, and sex of the baby. *New England Journal of Medicine*, *333*(23), 1517–1521.