Exploiting data missingness in Bayesian network modeling

Sérgio Rodrigues de Morais, Alexandre Aussem

Exploiting data missingness in Bayesian network modeling

Sérgio Rodrigues de Morais* — **Alexandre Aussem****

LIESP - University of Lyon
UCBL, 69622, Villeurbanne, France
* *sergio.rodrigues-de-morais@insa-lyon.fr*
** *aaussem@univ-lyon1.fr*

RÉSUMÉ. *Cet article présente une approche basée sur les réseaux Bayésiens pour représenter les dépendances entre les variables aléatoires et des variables indicatrices qui représentent la présence ou l'absence des valeurs des variables aléatoires auxquelles est sont associées. L'apprentissage de la structure du réseau augmenté permet, dans certains cas, d'indentifier le mécanisme par lequel les données sont manquantes. La méthode est illustrée sur des données synthétiques et sur un cas réel.*

ABSTRACT. *This paper proposes a framework built on the use of Bayesian networks (BN) for representing statistical dependencies between the existing random variables and additional dummy boolean variables, which represent the presence/absence of the respective random variable value. We show how augmenting the BN with these additional variables helps pinpoint the mechanism through which missing data contributes to the classification task. The missing data mechanism is thus explicitly taken into account to predict the class variable using the data at hand. Experiments on synthetic and real-world incomplete data sets are reported.*

MOTS-CLÉS : *Réseaux Bayésiens, mécanisme des données manquantes.*

KEYWORDS: *Bayesian networks, missing data mechanism.*

e soumission à *Journées Francophones sur les Réseaux bayésiens*, le 26 Mars 2010

1. Introduction

According to (Rubin, 1976), the assumptions about the missing data mechanisms may be classified into three categories : 1) missing completely at random (MCAR) : the probability that an entry is missing is independent of both observed and unobserved values in the data set ; 2) missing at random (MAR) : the probability that an entry is missing is a function of the observed values in the data set ; 3) informatively missing (IM) or Non-MAR (NMAR) : the probability that an entry is missing depends on both observed and unobserved values in the data set. The methods for coping with missing values can be grouped into three main categories (Little *et al.*, 2002) : inference restricted to complete data, imputation-based approaches, and likelihood-based approaches. Unfortunately, these methods are based on the assumption that the mechanism of missing data is not IM. This assumption is hard to test in practice (Statistical tests have been proposed, but these are restricted to a certain class of problems) and the decrease in accuracy may be severe when the assumption is violated. For instance, when machine learning algorithms are applied to data collected during the course of clinical care, the absence of expected data elements is common and the mechanism through which a data element is missing often involves the clinical relevance of that data element in a specific patient (Lin *et al.*, 2008; Siddique *et al.*, 2008). Hence the need for methods that help to detect the censoring mechanism. While no method can tell for sure, under all scenarios, from the data alone whether the missing observations are IM (although it is possible to distinguish between MCAR and MAR), some mechanisms leading to missing data actually possess information and the missingness of some variables can be a predictive information about other variables.

Recently, (Lin *et al.*, 2008) experimented with a method of treating missing values in a clinical data set by explicitly modeling the absence of data. They showed that in most cases a Naive Bayesian network trained using the explicit missing value treatments performed better. However there method is unable to pinpoint explicitly the missing mechanism and their experiments focus on small clinical datasets and thus the results may not generalize to other settings. Note also that several approaches have been designed with a view to be 'robust' to the missing data mechanism (Ramoni *et al.*, 2001; Aussem *et al.*, 2010). No assumption about the unknown censoring mechanism is made, hence the "robustness". However, the utility of these methods is questionable when the percentage of missing data is high. In this study, we experiment a new graphical method of treating missing values, based on Bayesian networks (BN). We describe a novel approach that uses explicitly the information represented by the absence of data to help detect the missing mechanism and reduce the classification error. We create an additional dummy boolean variable R_j to represent missingness for each existing variable X_j that was found to be absent (missingness indicator approach). The graphical structure of the BN representing the joint probability distribution of the variables can be used to help identify the missingness mechanism. Our approach is based on Markov boundary (MB for short) learning techniques to impute the missing entries. Once all the missing data are imputed, visual inspection of the induced graph reveals useful information on the the missing data mechanism. Several experiments

on synthetic incomplete data sets are reported. We also illustrate the usefulness of this approach to Nasopharyngeal Carcinoma (NPC) data. The data set is obtained from a case-control epidemiologic study performed by the International Agency for Research on Cancer in the Maghreb (north Africa).

2. Preliminaries

The missing data mechanisms can be graphically represented by meanings of a the structure of a Bayesian network, as stated by the following definition :

Definition 1 *Let $\mathcal{G}$ be a network structure, let $\mathcal{D}$ be a corresponding data set, and let $\boldsymbol{M}$ be the variables of $\mathcal{G}$ that have missing values in the dataset. Let $\boldsymbol{R}$ be the set of variables called **missing-data indicators** that are in one-to-one correspondence with variables $\boldsymbol{M}$. A network structure that results from adding variables $\boldsymbol{R}$ as leaf nodes to $\mathcal{G}$ is said to explicate the **missing-data mechanism** and is denoted by $\mathcal{G}_{\boldsymbol{R}}$.*

The following definition 2 provides a simple condition for the MCAR missing data mechanism :

Definition 2 *Let $\mathcal{G}_{\boldsymbol{R}}$ be a network structure that explicates the missing-data mechanism of structure $\mathcal{G}$ and data set $\mathcal{D}$. Let $\boldsymbol{O}$ be the set of variables that are always observed in the data set $\mathcal{D}$ and let $\boldsymbol{M}$ be the variables that have missing values in the data set. We say that $\mathcal{G}_{\boldsymbol{R}}$ satisfies the **missing completely at random** (MCAR) assumption if $\boldsymbol{R}$ and $\{\boldsymbol{M}, \boldsymbol{O}\}$ are disconnected in structure $\mathcal{G}_{\boldsymbol{R}}$.*

In other words, data will be MCAR if all the variables **R** form a sub-network that is disconnected from the other variables. Informal definitions of MAR and IM, as given in the introduction, only convey the general nature of the MAR and IM assumtpions. We need to ground our work in proper definitions as found in (Rubin, 1976; Darwiche, 2009) to make sense of the statement "the probability that an entry is missing is a function of the observed values". The next definition 3 provides a general condition for the MAR missing data mechanism (Darwiche, 2009) :

Definition 3 *We say that $\mathcal{G}_{\boldsymbol{R}}$ satisfies the **missing at random** (MAR) assumption if $\boldsymbol{R}$ and $\boldsymbol{M}$ are d-separated by $\boldsymbol{O}$ in structure $\mathcal{G}_{\boldsymbol{R}}$.*

Intuitively, $\mathcal{G}_{\mathbf{R}}$ satisfies the MAR assumption if once we know the values of the variables **O** the specific values of variables **M** become irrelevant to whether these values are missing in the data set. If the MAR assumption holds, the missing data mechanism can be ignored as shown by the followong theorem (Darwiche, 2009) :

Theorem 1 *Let $\mathcal{G}_{\boldsymbol{R}}$ and $\mathcal{D}_{\boldsymbol{R}}$ be a structure and a data set that explicate the missing-data mechanism of $\mathcal{G}$ and $\mathcal{D}$. Let θ be the parameters of structure $\mathcal{G}$ and $\theta_{\boldsymbol{R}}$ be the*

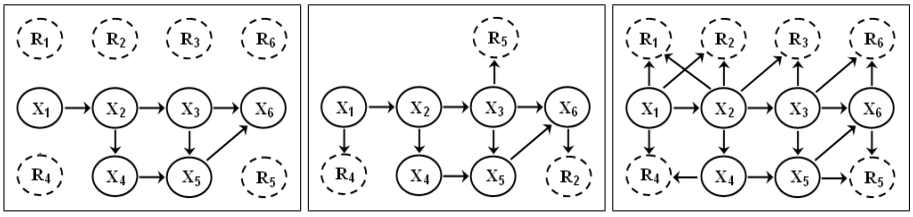

Figure 1. Graphical representation of the missing data mechanisms MCAR, MAR and NMAR (IM), respectively.

parameters of indicator variables in structure $\mathcal{G}_{\boldsymbol{R}}$. If $\mathcal{G}_{\boldsymbol{R}}$ satisfies the MAR assumption, then

$$argmax_{\theta} LL(\theta|\mathcal{D}) = argmax_{\theta} max_{\theta_{\boldsymbol{R}}} LL(\theta, \theta_{\boldsymbol{R}}|\mathcal{D}_{\boldsymbol{R}})$$

The data is IM if it not MAR. Note that the **R** are not necessary leaf nodes. Consider for example the network $R \rightarrow O \leftarrow M$. R and M are dependent given the oberved variable O, therefore the data is not MAR (noted IM or NMAR). To illustrate all these concepts, let us consider the toy examples presented in Figure 1. Those examples were originally used in (Ramoni *et al.*, 2001). Whittaker reports a dataset (Whittaker, 1990) that involves six boolean risk factors $X_1, \ldots, X_6$ observed in a sample of 1841 employees of a Czech car factory. Then Ramoni and Sebastiani considered that dataset and used a structure learning algorithm to output a structure that they used afterwards as a toy problem to learn the conditional probability tables from incomplete datasets (Ramoni *et al.*, 2001). Clearly the network in the left-side is an example of the mechanism MCAR, whilst the mechanism of missing data present in the network in the middle is an instance of MAR. Finally, the network in the right-side is a case of the mechanism NMAR because the missing-data indicators R_i depend on both observed and unobserved values in the dataset produced from this network.

Now let us consider the original data set $\mathcal{D}$ and an extension $\mathcal{D} \subset \mathcal{D}_{\mathbf{R}}$ that includes missing-data indicators. When we apply a machine learning algorithm on the data set $\mathcal{D}$ we are ignoring the missing-data mechanism, whilst we are accounting for it when we apply the algorithm on $\mathcal{D}_{\mathbf{R}}$. It turns out that the first and second approaches indeed yield different estimates of the parameters of a structure $\mathcal{G}$ when data are not missing at random.

From Definitions 2 and 3, the detection of the missing data mechanism boils down to analysing the topology of the augmented network $\mathcal{G}_{\mathbf{R}}$. In the MCAR case, a BN structure learning algorithm is able to infer that **R** and $\{\mathbf{M}, \mathbf{O}\}$ are disconnected. The same is true for the mechanism MAR where the structure learning algorithm should be able, in certain cases, to detect the presence of the edge between (at least) two variables R and O as both variables are observed.

A different picture emerges when one examines the NMAR case. If $R \rightarrow M$ in the generating graph, it is clear that the edge will not be inferred because, each time M is not missing, R is equal to 1. As R is constant, R and M appear as independent. To infer the dependency, the missing values for M should be imputed given exogeneous information. This should be possible if, for instance, $R \rightarrow M \leftarrow O$. In this case, O provides some information to infer the missing values for M. Once M is imputed, the link between R and M might be inferred.

Obverall, theses examples suggest that encountering a situation where the dataset has missing values should not discourage the researcher from applying a statistic principled method. Rather, the attitude should be to account for as much of the missing data mechanism as possible, knowing that these results will likely be better than those produced by methods which do not consider such mechanism during the learning process. Moreover, the missing data mechanism is rarely completely inaccessible. Often, the mechanism is actually made up of both accessible and inaccessible factors. Thus, although a researcher may not be confident that the data present a purely accessible mechanism, covering as much of the mechanism as possible should be regarded as beneficial rather than detrimental.

3. Imputation and detection

Our learning approach is rather heuristic in nature. We treat each variable that contains missing values as a variable to be imputed, then we fill each missing value with a single estimate (single imputation). When all vaiables are imputed, the network structure reveals some information on the data missing mechanism. In order to generate imputations for the missing values, one must impose a probability model on the complete data (observed and missing values). This is where the inferred BN model comes into the picture. The proposed imputation approach works in phases. First, a set of relevant variables is searched for building a probability model of the missing variable, and second, a BN is constructed on this set of variables. The problem of finding relevant predictive features is achieved in the context of determining the Markov boundary of the class variable that we want to predict. However, as some of these variables may have missing values as well, the idea is to induce a broader set of features that would not be strictly relevant for classification purposes if the dataset was complete, but that are still associated to the target. Therefore, the MB learning algorithm is called several times recursively to construct a local BN around the target. We call *MBLearning* the generic procedure applied for seeking the Markov boundary of a target from a dataset. This procedure can be replaced by any of the current state-of-the-art Markov boundary searching algorithms, such as those described in (Peña *et al.*, 2007; Rodrigues de Morais *et al.*, 2010; Tsamardinos *et al.*, 2006). The local graph provides a broader picture of the features that carry some information about the target variable. If the dataset was complete, these additional variables would deteriorate classification accuracy due to increasing design cost. This is not the case here as the variables in the true Markov boundary may be missing. A second important characte-

ristic of the method presented in this section is that the scope of the search process is augmented by the **R** variables.

Algorithm 1 *GMB*

Require: T : target variable ; r : maximal number of iterations for FSS ; α : minimal considered ratio for missing values ; $\mathcal{D}$: data set.
Ensure: **BN** : Bayesian network.

1: $\mathbf{U} = (X_{i..n} \cup R_{i..n})$
2: **Set1** $\leftarrow$ ***MBLearning***$(T, \mathbf{U})$
3: **V** $\leftarrow$ **Set1** $\cup\, T$
4: $I \leftarrow 1$
5: **while** $I < r$ **do**
6: **Set2** $\leftarrow \emptyset$
7: **for all** ($X_i \in$ **Set1**, such that ***MissRatio***$(X_i) \geq \alpha$) **do**
8: **Set2** $\leftarrow$ **Set2** $\cup$ ***MBLearning***$(X_i, \mathbf{U})$
9: **end for**
10: **Set1** $\leftarrow$ **Set2**
11: **V** $\leftarrow$ **V** $\cup$ **Set2**
12: $I \leftarrow I + 1$
13: **end while**

14: **BN** $\leftarrow$ Build Bayesian Network for **V**

The iterative algorithm is called *Growing Markov Boundary* (*GMB* for short). *GMB* (described in Algorithm 1) receives four parameters : the target variable (T), the maximal number of iterations (r), the minimal considered ratio for missing values (α) and the data set ($\mathcal{D}$). *GMB* proceeds as follows : first the scope of variables (**U**) is created in line 1 of the algorithm. **U** is composed by all the original random variables (X_i) and artificially created variables (R_i) representing missingness of their respective random variables. *MBLearning* is then first run on the target variable T (line 2). It is then run again repeatedly on the adjacent nodes and so on up to a radius of r around the target node (lines 5-13). A similar approach was proposed in (Peña *et al.*, 2005), but it does not take into account missingness as a possible piece of information. After finishing the feature subset selection process, *GMB* creates at line 14 the local BN including the selection of existing and the dummy variables. The user-defined radius (r) of the Bayesian network constructed by *GMB* trades off accuracy and scalability. *MissRatio(X)* is the missing rate of X.

It is important to note that *MBLearning* builds the MB in the presence of missing data. The structural learning can be performed with EM or MCMC techniques for instance. In this study, we adopt the simple *available cases analysis* (ACA), i.e., the contingency table for the test $X \perp_P Y|\mathbf{Z}$ is constructed on the cases having no missing values for X, Y and $\forall i, Z_i \in \mathbf{Z}$. Of course, this will bias results if the remaining cases are not representative of the entire sample. Notice however that ACA was empirically compared to EM in (François, 2006) leading to the conclusions that Bayesian network structure learning methods using ACA are faster and do not loose

accuracy compared to the same methods using EM. The algorithm *MBOR* (Rodrigues de Morais *et al.*, 2010) was used to implement *MBLearning*.

The shortcomings of signe-imputation have been well documented (Rubin, 1976). Without special corrective measures, single-imputation inference (contrary to multiple-imputation) tends to overstate precision because it omits the between-imputation component of variability. However if the proportion of missing values is small, then single imputation may be quite reasonable. We assume here that our single-imputation inferences for scalar estimands are fairly accurate although the statistical uncertainty due to missing data is not captured. Of course, our model used to generate the imputations will at best be only approximately true. Future work will generalize this appraoch to multiple-imputation.

Figure 2 reports the results of an empirical evaluation of *GMB* for MCAR, MAR and IM (NMAR). In this experiments we assess how the use of explicit representation of missing data affects classification across a range of different amounts of missing values, sample size and missing data mechanisms. We caused about 5%, 15% and 25% of the values to be missing according to MCAR, MAR and NMAR mechanisms by modifying the probability tables of the toy BN presented in Figure 1. Tow sample sizes (i.e., 500 and 1500) are considered in the experiments. X_6 was considered as target because this variable permits a maximal value for the parameter r. We run *GMB*(X_6) for $r = 1, 2, 3$ and the local BN output by *GMB* was used as the classifier for X_6 using standard inference techniques. Figure 2 summarizes the variability of the Kappa measure by 10-fold cross-validation. The Kappa distribution over 50 datasets is illustrated in the form of boxplots. The Kappa measure assesses improvement over chance. The following ranges of agreement for the Kappa statistic suggested in the literature are : poor $K < 0.4$, good $0.4 < K < 0.75$ and excellent $K > 0.75$. As may be seen in Figure 2, the prediction value derived from missing data appears to be useful for increasing the accuracy of the toy problem when the percentage of missing data is superior to 5%. The term 'MB' denotes the classifier using only the MB of the target variable without the use of the dummy variables R_i. The analysis presented here suggests that attention to missing data may improve the prediction accuracy. Further conclusions can be drawn from these results. In the MCAR case, the inclusion of the dummy variables cannot improve classification because they are independent of all the variables. The observed improvement for $r > 1$ is only due to the additional X_i variables that are found useful when others are missing. A radius $r > 1$ was not found to improve significantly the classification, compared to $r = 1$, when data are missing by MAR or NMAR. The usefulness of the dummy variables increases with the ratio of missing data when data are MAR or NMAR. Finally, the size of the dataset has little influence on the results when data is MAR and NMAR.

4. Nasopharyngeal carcinoma analysis

The NPC data set consists of 1289 subjects (664 cases of NPC and 625 controls), 150 nominal variables and 5% missing values. Here the local structure around the tar-

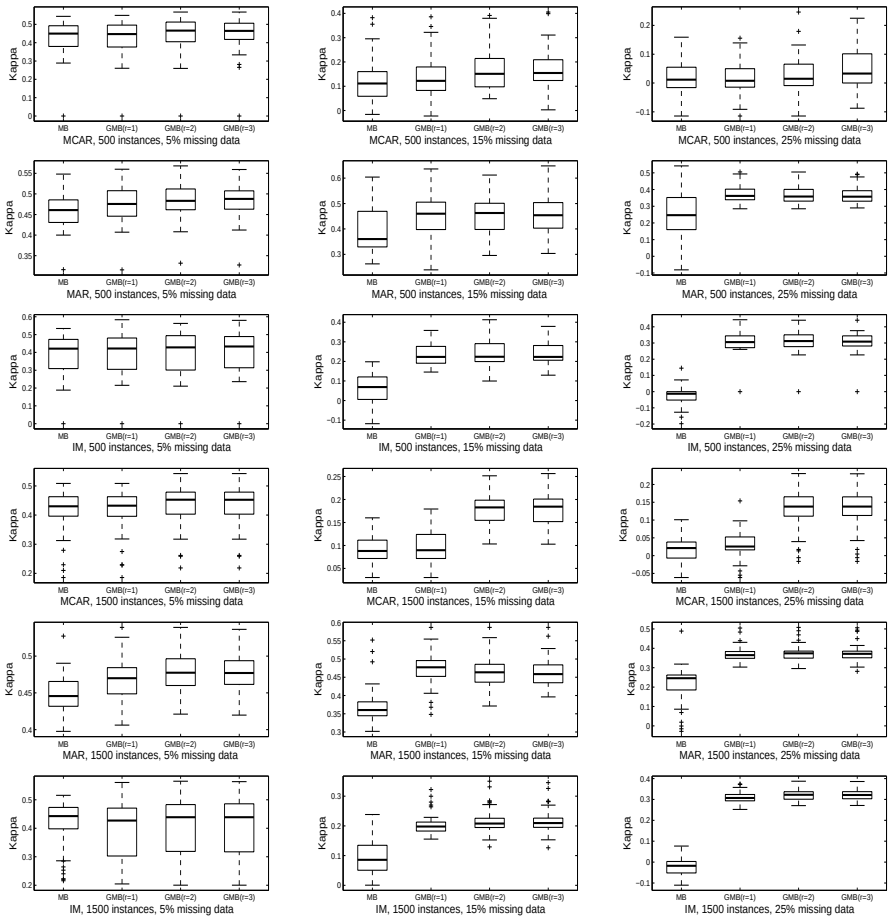

Figure 2. Accuracy of GMB for MCAR, MAR and IM missing data mechanisms on synthetic data, for 5%, 15% and 25% missing values, and 500, and 1500 instances .

get variable NPC was obtained by running *GMB*. *GMB* was run for $r = 1$ on each existing variable X_j with missing values and we used the BN output by *GMB* as the classifier using standard inference techniques to fill in the missing entries for X_j. Once all missing values were imputed, the overall structure was constructed by applying *HPC* on the completed dataset. Figure 3 shows the local graph obtained. The nodes in dotted lines correspond to the dummy missingness variables. Line width is proportional to the G-statistic association measure discussed earlier. The emphasis here is on the missingness information. As may be observed, Nonresponse to "cigarette smoking duration" and "stop smoking age" are directly associated to NPC even if smoking is known to have a marginal effect on NPC. In fact, smoking is known to be highly correlated to lifestyle habits in the maghrebian societies but not to NPC as

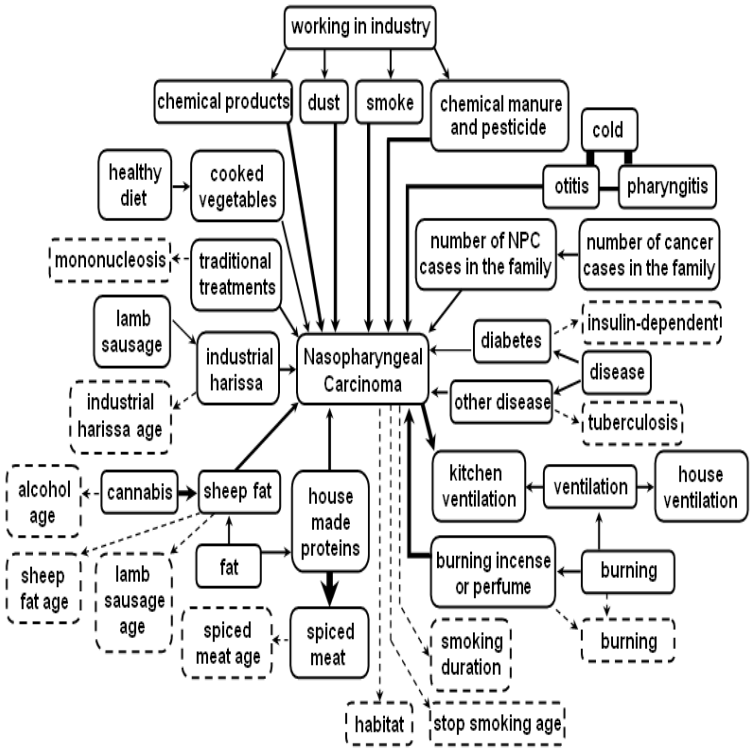

Figure 3. Local NPC graph with missingness variables shown in dotted line.

NPC is less sensitive to the carcinogenic effects of tobacco constituents. It is worthy to note that these variables are considered by our expert domain as effects of NPC instead of causes because he believes that NPC patients are more inclined to answer these questions as they are anxious about the effects of smoke inhalation. Inspection of the graph also reveals that "industrial harissa age", "sheep fat", "lamb sausage age", "spiced meat age", "tuberculosis", "alcohol age", "insulin-dependent" and "mononucleosis" appear to be MAR and that the nonresponse to these questions is indirectly associated to NPC (their influence is mediated by other variables and they are not in the Markov boundary of NPC). Moreover, the missingness of "burning" seems to be NMAR. A possible reason is that individuals skip this item if they are not exposed to smoke particles from incomplete combustion of coal and wood. Finally, inclusion of the missingness variables seems here to increase a little the prediction accuracy (about 2%). While the gain in prediction accuracy is relatively moderate due the limited amount of missing data (5%), the present study confirms that non response to smoking and habitat conditions are associated to NPC risk.

5. Discussion and conclusions

In study, we discussed a model for the imputation and the detection of the missing data mechanism. Although absence of data is usually considered a hindrance to accurate prediction, our conclusion is that the absence of some data elements in the data sets can be informative when the amount of missing data is greater than 5%. Further experiments were reported in (Rodrigues de Morais *et al.*, 2009). Future work will extend this approach to multiple imputation.

6. Bibliographie

Aussem A., Rodrigues de Morais S., « A Conservative Feature Subset Selection Algorithm with Missing Data », *Neurocomputing*, vol. 73, p. 585-590, 2010.

Darwiche A., *Modeling and Reasoning with Bayesian Networks*, Cambridge University Press, New York, 2009.

François O., *De l'indentification de la structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes*, Thèse de doctorat de l'INSA Rouen, 2006.

Lin J., Haug P., « Exploiting missing clinical data in Bayesian network modeling for predicting medical problems », *Journal of Biomedical Informatics*, vol. 41, p. 1-14, 2008.

Little R., Rubin D., *Statistical analysis with missing data*, Whiley-Interscience, 2002.

Peña J. M., Björkegren J., Tegnér J., « Growing Bayesian Network Models of Gene Networks from Seed Genes », *Bioinformatics*, vol. 40, p. 224-229, 2005.

Peña J., Nilsson R., Björkegren J., Tegnér J., « Towards Scalable and Data Efficient Learning of Markov Boundaries », *International Journal of Approximate Reasoning*, vol. 45, n° 2, p. 211-232, 2007.

Ramoni M., Sebastiani P., « Robust Learning with Missing Data. », *Machine Learning*, vol. 45, n° 2, p. 147-170, 2001.

Rodrigues de Morais S., Aussem A., « Exploiting data missingness in Bayesian network modeling », *8th International Symposium on Intelligent Data Analysis, Lyon, France*, vol. 5772 of *Lecture Notes in Computer Science*, Springer, p. 35-46, 2009.

Rodrigues de Morais S., Aussem A., « A Novel Markov Boundary Based Feature Subset Selection Algorithm », *Neurocomputing*, vol. 73, p. 578-584, 2010.

Rubin D., « Inference and missing data », *Biometrika*, vol. 63, p. 581-592, 1976.

Siddique J., Belin T., « Using an Approximate Bayesian Bootstrap to multiply impute nonignorable missing data », *Computational Statistics & Date Analysis*, vol. 53, p. 405-415, 2008.

Tsamardinos I., Brown L., Aliferis C., « The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. », *Machine Learning*, vol. 65, n° 1, p. 31-78, 2006.

Whittaker J., *Graphical Models in Applied Multivariate Analysis*, New York. Whiley, 1990.