Implicit parameter estimation for conditional Gaussian Bayesian networks

Aida Jarraya[1,2], **Philippe Leray**[2], **Afif Masmoudi**[1]

[1] *Laboratory of Probability and Statistics*
Faculty of Sciences of Sfax, Sfax University. B. P. 1171 Sfax. Tunisia
E-mail: jarraya_aida@yahoo.fr, Afif.Masmoudi@fss.rnu.tn
[2] *LINA Computer Science Lab UMR 6241*
Knowledge and Decision Team, University of Nantes, France.
E-mail: philippe.leray@univ-nantes.fr

Received 12 June 2012

Accepted 7 August 2013

Abstract

The Bayesian estimation of the conditional Gaussian parameter needs to define several a priori parameters. The proposed approach is free from this definition of priors. We use the Implicit estimation method for learning from observations without a prior knowledge. We illustrate the interest of such an estimation method by giving first the Bayesian Expectation A Posteriori estimator for conditional Gaussian parameters. Then, we describe the Implicit estimators for the same parameters. Moreover, an experimental study is proposed in order to compare both approaches.

Keywords: Conditional Gaussian Bayesian networks; Bayesian estimation; Implicit estimation; Parameter learning.

1. Introduction

Bayesian Networks (BNs) are probabilistic graphical models widely used for knowledge representation and reasoning within an uncertain framework [1,2,3]. Learning Bayesian networks from data, i.e., obtaining automatically the structure and parameters from information belonging to the available samples, is a NP hard problem [4]. In this paper, we focus on the first component of BN learning which is parameter learning. Classical methods such as Maximum Likelihood (ML) or Bayesian methods such as Maximum A Posteriori (MAP) or Expectation A Posteriori (EAP) can be used for parameter learning, whatever the BN parametrization: discrete BNs, linear Gaussian BNs, or conditional Gaussian BNs.

In fact, by using the Bayesian approach, the posterior distribution is given by multiplying the likelihood function by a known prior distribution and then dividing by a norming constant. Consequently, this prior information is used, together with the data, in order to derive the posterior distribution. The use of prior allows us to take into account the expert knowledge. Nevertheless, this prior is not always available. As we know, the choice of a specific prior information in Bayesian approaches is often problematic and is even considered to represent the major weakness of such methods because a biased result can be obtained if we make a bad choice of the prior distribution. Hence, if we can find the posterior distribution with the data likelihood, the method will be much easier to use. This represents the principle of the Implicit approach [5]. This approach is similar to the Bayesian one, and happens

Co-published by Atlantis Press and Taylor & Francis

in a natural context without specifying any prior parameters. We use the Implicit method in order to overcome some of the shortcomings associated with the Bayesian estimation concerning the choice of the prior distribution parameters. The implicit approach has already been applied for learning parameters in discrete BNs [6,7] with an extension for structure learning [8]. We propose here new theoretical results concerning Implicit learning in Conditional Gaussian Bayesian Networks (CGBNs) in different contexts used in real applications by dealing with tied or untied parameters. In such models, priors implied can be more complex to express as compared with the usual Dirichlet priors for discrete variables. By going back to the classical notations used in the domain, and following Murphy's work [9] devoted to Maximum Likelihood (ML) and Bayesian Maximum A Posteriori (MAP) estimation for conditional Gaussian models, we enlarge them with Expectation A Posteriori (EAP) and Implicit approaches. The outline of this paper is organized as follows: in section 2, we briefly present the Implicit method and recall the principles of Implicit estimation. In section 3, we discuss the problem of parameter learning in CGBNs by using Bayesian estimation. In section 4, we present parameter estimation in CGBNs using Implicit approach. Then, both approaches are formally compared in section 5 and experimental results are provided in section 6. Finally, we conclude with perspectives for future work.

2. Implicit estimation

2.1. Principle

The Bayesian estimation method gathers the information of the data (the data likelihood) with the information collected from past experience (prior distribution) and finds a new updated information (posterior distribution). The Implicit estimation is an alternative to Bayesian estimation and does not need to specify any prior for the parameters. In the context of the Bayesian theory, the unknown parameter θ in a statistical model is assumed to be a random variable with a known prior distribution. This prior information is used, together with the data, in order to derive the posterior distribution of θ. The choice of a prior is generally based on the preliminary knowledge of the problem. So, the basic idea of the Bayesian theory is to consider any parameter θ as a random variable and to determine its posterior distribution given the data and the assumed prior.
Alternatively, the concept of Implicit distribution previously proposed by[5], can be described as a kind of posterior distribution of a parameter given the data. To explain the principle of Implicit distribution, let us consider a family of probability distributions $\{p(x|\theta), \theta \in \Theta\}$ parameterized by an unknown parameter θ in a set Θ; where x represents the observed data.
The Implicit distribution is computed by multiplying the likelihood function $p(x|\theta)$ by a counting measure σ if Θ is a countable set and by a Lebesgue measure σ if Θ is an open set (σ depends only on the topological structure of Θ) and then, dividing by the norming constant $c(x) = \int_\Theta p(x|\theta)\sigma(d\theta)$. Therefore, the Implicit distribution is given by the following formula

$$Q_x(d\theta) = \{c(x)\}^{-1} p(x|\theta)\sigma(d\theta)$$

and plays the role of a posterior distribution of θ given x in the Bayesian method, corresponding to a particular improper prior which depends only on the topology of Θ (without any statistical assumption). Provided its existence (which holds for most statistical models), the Implicit distribution can be used for the estimation of the parameter θ following a Bayesian methodology. The Implicit estimator $\widehat{\theta}$ of θ is nothing but the mean of the Implicit distribution, that is

$$\widehat{\theta} = E(\theta|x) = \int_\Theta \theta Q_x(d\theta).$$

Readers are referred to the paper from [5] for a presentation of the theoretical foundations of Implicit estimation and some selected applications.

2.2. Example with variance estimation

Let $X_i \sim N(0, \sigma^2)$ be a centered Gaussian random variable with an unknown variance $\sigma^2 \in]0, +\infty[$, the likelihood of σ^2 for n independent observations $\underline{x} = (x_1, \ldots, x_n)$ is $l(\underline{x}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\}$.

and then, its sample Implicit distribution $Q_{\underline{x}}(d\sigma^2)$ is given by

$$Q_{\underline{x}}(d\sigma^2) = \frac{(\sigma^2)^{-\frac{n}{2}} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\}}{\Gamma(\frac{n}{2}-1)} (\frac{1}{2}\sum_{i=1}^{n} x_i^2)^{\frac{n}{2}-1}.$$

By a standard calculation, we show that the Implicit estimator of σ^2 is

$$(\widehat{\sigma}^2)^{Imp} = E(\sigma^2|X_1,\ldots,X_n) = \frac{1}{n-4}\sum_{i=1}^{n} X_i^2 \qquad (1)$$

with $n > 4$, which is different from the ML estimator for which the normalizing factor is $\frac{1}{n}$. We can compare this result with the Bayesian estimation obtained by the EAP (Expectation A Posteriori) approach

$$(\widehat{\sigma}^2)^{Bay} = \frac{\sum_{i=1}^{n} X_i^2 + 2b}{n+2a-2}, \qquad (2)$$

where the prior distribution of σ^2 is an Inverse-Gamma $IG(a,b)$ with a shape parameter a and a scale parameter b.

First, we can see that the Implicit estimator does not need to tune any hyper-parameters such as a and b in the EAP approach. We can also notice that the equality between the two estimators of variance obtained by Bayesian and Implicit approaches is established for $a = -1$ and $b = 0$, which is impossible since a and b (parameters of an Inverse Gamma distribution) have to be positive. This shows that the Implicit estimator does not correspond to a Bayesian one for specific prior values. To compare the performances of Implicit and Bayesian method, we start (as a first step) by simulation of data using Matlab software, and (in the second step) we validate our results by comparing them with those of the Bayesian method. To compare two statistical approaches and to appreciate in which measure the result will be more definite, we choose an indicator which is the Mean Squared Errors (MSE) between the estimator and the true value of the parameter. We generated 1000 observations from the Gaussian model, then we compare the Bayesian and Implicit estimators in terms of the mean squared errors (MSE) for different true parameter values of σ^2. We replicate the process 10000 times and we compute the average estimates and the MSE. We perform two Bayesian simulation studies based on two different prior densities for the parameter σ^2. The results are reported in Table 1.

Bayesian estimation (Bay*) is obtained by specifying true values as prior parameters. Bayesian estimation (Bay**) corresponds to parameters estimated by using prior values different from the true ones.

The comparison of MSE obtained by Bay* and Bay** proves the sensibility of the parameter estimation with respect to the choice of the prior distribution. This result proves the fragility of the Bayesian estimation. A biased result can be obtained if we make a bad choice of the prior distribution. Results obtained by Implicit approach (without need of any prior information) and Bay* approach (based on true values as priors) are close to the true values. We notice a very good concordance between both approaches. However, we may point out a better precision of parameter estimated by Implicit method than the corresponding one for Bay** method (with priors different from true values). These results are illustrated in Fig.1. The yellow part corresponds to the area where the Implicit MSE is smaller than the Bayesian MSE with respect to the values of the prior coefficients a and b. Whatever the value of σ^2 (a low or a high one), the Implicit MSE is often lower than the Bayesian one.

2.3. Related works

As seen previously, the Implicit estimation does not rely on a prior definition. Another alternative to Bayesian method, with the objective to get a distribution of the unknown parameter θ without any priors is the Fiducial distribution introduced by [10]. It describes the uncertainty about the value of the fixed unknown parameter θ by supposing that there is a population characterized by the density function $p(x,\theta)$, where the form of the density $p(x,\theta)$ is known, but there is no prior information available about the true value of the parameter θ [11].

Mukhopadhyay [12] claimed that the Implicit inference is nothing new and that it is either a Fiducial-like approach or a non-informative Bayesian method. Concerning his criticism, our

$\overline{\sigma^2}$	priors	$(\widehat{\sigma^2})^{Imp}$	MSE(Imp)	$(\widehat{\sigma^2})^{Bay^*}$	$MSE(Bay^*)$	$(\widehat{\sigma^2})^{Bay^{**}}$	$MSE(Bay^{**})$
0.1	a=2, b=0.1	0.1	$2\ .\ 10^{-6}$	0.1	$2\ .\ 10^{-6}$	0.11	$1.03.\ 10^{-5}$
0.05	a=2, b=0.05	0.05	$4.99.\ 10^{-7}$	0.05	$5.\ 10^{-7}$	0.053	$9.39.\ 10^{-6}$
1	a=2, b=1	1.0002	$1.99.\ 10^{-4}$	1.0006	$1.99.\ 10^{-4}$	1.019	$5.87.\ 10^{-4}$
1.5	a=2, b=1.5	1.5	$4.45.\ 10^{-4}$	1.5006	$4.46.\ 10^{-4}$	1.52	0.0013
2	a=2, b=2	2	$8.03.\ 10^{-4}$	2.0007	$8.05.\ 10^{-4}$	2.039	0.0024

Table 1: Estimators of σ^2 obtained by Implicit method, Bayesian method with true a priori (Bay^*) and different one (Bay^{**}), for several prior parameters (a and b) and σ^2.

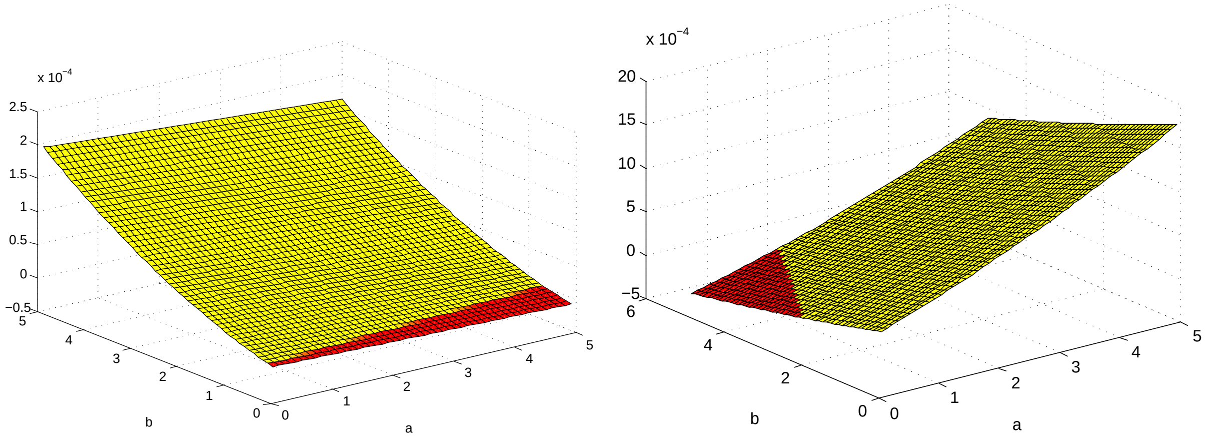

Figure 1: Difference between Bayesian and Implicit mean squared errors with respect to prior parameters a and b for $\sigma^2 = 0.1$ and $\sigma^2 = 2$.

comment is to show that Implicit inference is in fact a new paradigm in statistical inference. Using several examples, we show that, in many cases when the parameter space θ is infinite, Implicit distribution does not coincide with neither Fiducial nor Bayesian distribution. If the parameter space θ is bounded, then the Implicit distribution coincides with the posterior distribution with uniform prior in Bayesian method. The coincidence of both Implicit and Fiducial distributions in the normal model $N(\theta,1)$ with a mean θ and a variance 1, seems to explain the misleading comments of [12]. In what follows, we give selected examples illustrating clearly the difference between Implicit, Bayesian and Fiducial approaches.

2.3.1. *Example 1: Binomial model $B(n,\theta)$*

In the Binomial case, applying the Implicit method gives:

$$c(x) = \frac{1}{n+1}.$$

It comes that the Implicit distribution of θ given x is a Beta distribution with parameters $x+1$ and $n-x+1$, denoted $Beta(x+1,n-x+1)$. Heike and al [11] showed that, for the same binomial model, the Fiducial distribution is a Beta distribution denoted $Beta(x,n-x+1)$, with parameters x and $n-x+1$.

2.3.2. *Example 2: Exponential model $\varepsilon(\theta)$*

Let $X_1,\dots,X_n$ be n independent random variables identically distributed according to the exponential distribution with parameter θ. It is well known that

$\sum_{i=1}^{n} X_i$ follows a gamma distribution denoted $\gamma(n,\theta)$, with parameters n and θ.
The norming constant is given by

$$c(x_1,\ldots,x_n) = \frac{n!}{(\sum_{i=1}^{n} x_i)^{n+1}}.$$

Then, the Implicit distribution of θ given $X_1,\ldots,X_n$ is a gamma distribution denoted $\gamma(n+1, \sum_{i=1}^{n} x_i)$, with parameters $n+1$ and $\sum_{i=1}^{n} x_i$. For the same model, [11] showed that the Fiducial distribution is a gamma distribution denoted $\gamma(n, \sum_{i=1}^{n} x_i)$, with parameters n and $\sum_{i=1}^{n} x_i$. Then, the difference between the two methods is very clear. The first parameter is $n+1$ in the case of Implicit method but, in the Fiducial method, it is n.

3. Parameter estimation in conditional Gaussian Bayesian networks using Bayesian approach

In this section, we formally define Conditional Gaussian Bayesian Networks. We inspire from Murphy's work [9] devoted to MAP and ML estimation for parameters of such models and we enlarge them with EAP estimation.

3.1. Definitions and notations

Bayesian Networks (BNs) are usually defined for discrete variables with a finite number of states. This assumption is not very realistic in several application areas such as medicine, where the elaboration of the diagnosis is generally the result of some mixture of information of continuous type (results of laboratory) and of discrete type (presence / absence of a symptom).

In the literature, previous works have concentrated on the study of probabilistic graphical models with both discrete and continuous variables [13,14,15]. In this paper, we are interested in domains containing either continuous variables or a mixture of both discrete and continuous variables, under the assumption that continuous data constitute a sample from a multivariate normal (Gaussian) distribution. Consider a finite set $X = \{X_1,\ldots,X_n\}$ of random variables. A Bayesian network (BN) is a directed acyclic graph G and a set of conditional probability distributions which represent a joint probability distribution[1]. The nodes of the graph correspond to the random variables and are annotated with a Conditional Probability Density (CPD) of the random variable given its parents Pa_i in the graph G. The joint distribution is the product over families (variable and its parents)

$$p(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i|Pa_i).$$

The graph G represents independence properties which are assumed to hold in the underlying distribution: each X_i is independent from its non-descendants given its parents Pa_i.
Unlike the case of discrete variables (when the variable X and some or all of its parents are real valued), there is no representation that can integrate all conditional densities. A common choice is the use of linear Gaussian conditional densities [16], where each variable is a linear function of its parents. When all the variables in a network have linear Gaussian conditional densities, the joint density over X is a multivariate Gaussian. In order to simplify future equations, we summarize below the notations initially proposed by Murphy[9]. We will consider the problem of finding estimators for the parameters of a conditional Gaussian variable Y with continuous parent X and discrete parent Q, i.e., $p(y|x,Q=i) = c|\Sigma_i|^{-\frac{1}{2}} exp(-\frac{1}{2}(y-B_i x-\mu_i)'\Sigma_i^{-1}(y-B_i x-\mu_i))$ where c is a constant and $|y| = d$. We assume that we have N iid training cases $\{e_t\}$ so the complete data log-likelihood is $\log\prod_{t=1}^{N}\prod_{i=1}^{|Q|} p(y_t|x_t,Q_t=i,e_t)^{q_t^i}$ where $q_t^i = 1$ if Q has the value i in the t'th complete case, and 0 otherwise. Since Q,X and Y may all be unobserved, the expected complete-data likelihood is defined by $\widetilde{p}(y|x,Q=i) = \exp(l)$ with $l = -\frac{1}{2}\sum_t E\Big(\sum_i q_t^i \log|\Sigma_i| + q_t^i(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t -$

$B_ix_t - \mu_i)|e_t\Big)$.

We can write $E(q_t^i x_t x_t'|e_t) = E(q_t^i|e_t)E(x_t x_t'|Q_t = i, e_t) = w_t^i E(XX')$ where the weights $w_t^i = p(Q = i|e_t)$ are posterior probabilities and $E_{ti}(XX')$ is a conditional second moment. We get $l = -\frac{1}{2}\sum_t \sum_i w_t^i \log|\Sigma_i| - \frac{1}{2}\sum_t \sum_i w_t^i E_{ti}\Big((y_t - B_ix_t - \mu_i)'\Sigma_i^{-1}(y_t - B_ix_t - \mu_i)|e_t\Big)$.

The following expected sufficient statistics are introduced in order to simplify future equations:

$$w_i = \sum_t w_t^i$$

$$S_{XX',i} = \sum_t w_t^i E_{ti}(XX')$$

$$S_{X,i} = \sum_t w_t^i E_{ti}(X)$$

$$S_{YY',i} = \sum_t w_t^i E_{ti}(YY')$$

$$S_{Y'Y,i} = \sum_t w_t^i E_{ti}(Y'Y)$$

$$S_{Y,i} = \sum_t w_t^i E_{ti}(Y)$$

$$S_{XY',i} = \sum_t w_t^i E_{ti}(XY')$$

$$S_{YX',i} = \sum_t w_t^i E_{ti}(YX').$$

Usually, two situations can be considered when dealing with the parameters of CGBNs models. The first one (untied parameters) corresponds to the general one, with the parameters defined in the previous definition. In this situation, if one continuous variable has continuous and discrete parents, we will have to deal with conditional Gaussian parameters (mean, covariance, regression coefficients) for each configuration of the discrete parents. The second situation (tied parameters) considers that the conditional Gaussian parameters are independent from the discrete parents. This assumption reduces the number of parameters which can be interesting when the number of data is limited.

Another way to decrease the number of parameters when estimating the covariance matrix is to consider a spherical covariance matrix, i.e. the constraint that $\Sigma_i = \sigma_i^2 I$ is isotropic.

The remaining of this section will be devoted to the proposition of Bayesian estimation (with Expectation A Posteriori method) of conditional Gaussian parameters (regression coefficients, mean, covariance) for all these situations (untied or tied parameters, full or spherical covariance matrix).

3.2. *Regression matrix estimation*

3.2.1. *Untied parameters*

$\widetilde{p}(y|B_i, x_t, \Sigma_i, \mu_i) \propto \exp(-\frac{1}{2}\sum_{t=1}^{N}(y_t - B_ix_t - \mu_i)'\Sigma_i^{-1}(y_t - B_ix_t - \mu_i))$.

We classically assume that the prior for B_i is a multivariate normal distribution with mean a and covariance matrix V.

Hence, the posterior $\widetilde{p}(B_i|y_t, x_t, \Sigma_i, \mu_i) = exp(-\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{t_i}(y_t - B_ix_t - \mu_i)'\Sigma_i^{-1}(y_t - B_ix_t - \mu_i))exp(-\frac{1}{2}(B_i - a)'V^{-1}(B_i - a))$ is also a multivariate normal distribution with a mean given by

$$\begin{aligned}\widehat{B_i}^{Bay} &= (\Sigma_i^{-1}(S_{YX',i} - \mu_i S_{X',i}) + aV^{-1}) \\ &\quad (\Sigma_i^{-1}S_{XX',i} + V^{-1})^{-1} \qquad (3)\end{aligned}$$

3.2.2. *Tied parameters*

For the tied case, the estimator of B_i becomes

$$\begin{aligned}\widehat{B}^{Bay} &= (\sum_i(\Sigma_i^{-1}(S_{YX',i} - \mu_i S_{X',i}) + aV^{-1})) \\ &\quad (\sum_i(\Sigma_i^{-1}S_{XX',i} + V^{-1}))^{-1}. \qquad (4)\end{aligned}$$

3.3. *Mean estimation*

3.3.1. *Untied parameters*

For the mean parameter μ_i, the prior density is a multivariate normal distribution with parameters (m, ψ) [16,3]

$\widetilde{P}(\mu_i/y_t, x_t, \Sigma_i, B_i) \propto exp(-\frac{1}{2}\sum_{t=1}^{N} w_t^i E_{t_i}(y_t - B_ix_t -$

$\mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i))exp(-\frac{1}{2}(\mu_i - m)'\psi^{-1}(\mu_i - m))$

After some calculations applied to the posterior density and using the following identity

$$\frac{\partial(XA+b)'C(XA+b)}{\partial X} = (C+C')(XA+b)A' \quad (5)$$

we get the following expression

$$\begin{aligned} \widehat{\mu}_i^{Bay} \;&=\; (\psi^{-1}m + \Sigma_i^{-1}(S_{Y,i} - B_i S_{X,i})) \\ &\quad (w_i\Sigma_i^{-1} + \psi^{-1})^{-1}. \end{aligned} \quad (6)$$

3.3.2. *Tied parameters*

For the tied case, we get

$$\begin{aligned} \widehat{\mu}^{Bay} \;&=\; (\sum_i(\psi^{-1}m + \Sigma_i^{-1}(S_{Y,i} - B_i S_{X,i}))) \\ &\quad (\sum_i(w_i\Sigma_i^{-1} + \psi^{-1}))^{-1}. \end{aligned} \quad (7)$$

3.4. *Estimating the regression matrix and the mean simultaneously*

Since the equation for B_i depends on μ and vice versa, if both of them have to be estimated, we must estimate them jointly. We can do this by appending μ_i as the last column to B_i in order to create D_i, and also appending a 1 to the last component of X in order to create Z. Then, we have

$$p(y|x, Q=i) = c|\Sigma_i|^{-\frac{1}{2}} exp(-\frac{1}{2}(y - D_i z)'\Sigma_i^{-1}(y - D_i z)).$$

By using the equation 5 with $\mu_i = 0$ and replacing $S_{XX',i}$ by $S_{ZZ'}$ and also $S_{YX',i}$ by $S_{YZ',i}$ we get

$$\widehat{D}_i^{Bay} = (\Sigma_i^{-1}S_{YZ',i} + aV^{-1})(\Sigma_i^{-1}S_{ZZ',i} + V^{-1})^{-1}. \quad (8)$$

The substitutions are

$$E_{t_i}ZZ' = E_{t_i}\begin{pmatrix} XX' & X \\ X' & 1 \end{pmatrix}$$

so,

$$S_{ZZ',i} = \begin{pmatrix} S_{XX',i} & S_{X,i} \\ S_{X',i} & w_i \end{pmatrix}$$

and

$$E_{t_i}YZ' = E_{t_i}\begin{pmatrix} YX' & Y \end{pmatrix}.$$

Then,

$$S_{YZ',i} = \begin{pmatrix} S_{YX',i} & S_Y \end{pmatrix}.$$

3.5. *Full Covariance matrix estimation*

3.5.1. *Untied parameters*

We classically assume that the prior for Σ_i is an Inverse-Wishart distribution with α degrees of freedom and a positive definite precision matrix V which implies that the posterior density for the parameter Σ_i is proportional to the following expression: $\widetilde{p}(\Sigma_i|y_t, x_t, B_i, \mu_i) \propto |\Sigma_i|^{-\frac{w_i+\alpha+d+1}{2}} \exp(-\frac{1}{2}tr(\Sigma_i^{-1}\sum_t\sum_i w_t^i E_{ti}(y_t - B_i x_t - \mu_i)(y_t - B_i x_t - \mu_i)' + V^{-1}))$.

Hence, the posterior $\widetilde{p}(\Sigma_i|B_i, y_t, x_t, \mu_i)$ is also an Inverse-Wishart distribution with $(w_i + \alpha)$ degrees of freedom and a positive definite precision matrix $\sum_t w_t^i E_{ti}(y_t - B_i x_t - \mu_i)(y_t - B_i x_t - \mu_i)' + V^{-1}$.

So,

$$\widehat{\Sigma}_i^{Bay} = \frac{1}{w_i + \alpha - d - 1}(A_i + V) \quad (9)$$

where $A_i = S_{YY',i} - S_{YX',i}B_i' - S_{Y,i}\mu_i' - B_i S_{XY',i} + B_i S_{XX',i}B_i' + B_i S_{X,i}\mu_i' - \mu_i S_{Y',i} + \mu_i S_{X',i}B_i' + \mu_i\mu_i'$.

If $B_i = 0$, then we have

$$\widehat{\Sigma}_i^{Bay} = \frac{S_{YY',i} - S_{Y,i}\mu_i' - \mu_i S_{Y',i} + \mu_i\mu_i' + V}{w_i + \alpha - d - 1}. \quad (10)$$

3.5.2. *Tied parameters*

if Σ_i is tied, we get

$$\widehat{\Sigma}^{Bay} = \frac{1}{N(\alpha - d)}\sum_i(A_i + V) \quad (11)$$

3.6. *Spherical covariance matrix estimation*

3.6.1. *Untied parameters*

$$p(y|x, Q=i) = c\sigma_i^{-d} exp(-\frac{1}{2}\sigma_i^{-2}\|y - B_i x_t - \mu_i\|^2).$$

For the σ_i^2, we classically assume that the prior is an Inverse-Gamma distribution with parameters (a, b) $\widetilde{p}(\sigma_i^2|y_t, x_t, B_i, \mu_i) \quad \propto \quad (\sigma_i^2)^{-\frac{dw_i}{2}-a-1}\exp\Big(\quad -$

$\frac{1}{2\sigma_i^2}(\sum_t w_t^i E_{ti}(\|y_t - B_i x_t - \mu_i\|^2) + b)\Big)$.
In fact, after some computations applied to the posterior density, we get the following expression for the estimator of σ_i^2

$$(\widehat{\sigma_i^2})^{Bay} = \frac{1}{dw_i + 2a - 2} tr(C_i + 2b) \qquad (12)$$

where
$C_i = S_{Y'Y,i} - B_i S_{Y'X,i} - S_{Y',i}\mu_i - B_i' S_{X'Y,i} + B_i' B_i S_{X'X,i} + B_i' \mu_i S_{X',i} - \mu_i' S_{Y,i} + \mu_i' B_i S_{X,i} + \mu_i' \mu_i$.
If we don't have any regression, then
$(\widehat{\sigma}_i^2)^{Bay} = \frac{tr(S_{Y'Y,i} - S_{Y',i}\mu_i - \mu_i' S_{Y,i} + \mu_i' \mu_i + 2b)}{dw_i + 2a - 2}$.

3.6.2. Tied parameters

If σ_i^2 is tied, we have

$$(\widehat{\sigma^2})^{Bay} = \frac{1}{N(d + 2a - 2)} tr \sum_i (C_i + 2b). \qquad (13)$$

4. Parameter estimation in Conditional Gaussian Bayesian Networks using Implicit approach

As seen in section 2.3, if the parameter space is bounded, then the Implicit distribution coincides with the posterior distribution with a uniform prior in Bayesian method, which is not the case for conditional Gaussian parameters.

Using the same notation of the previous section, we propose here the estimation of parameters by the Implicit method.

4.1. Regression matrix estimation

4.1.1. Untied parameters

Let $\widetilde{p}(y|x, Q = i) = \exp(l)$ $\widetilde{p}(B_i|y_t, x_t, \Sigma_i, \mu_i) \propto \exp(-\frac{1}{2}\sum_t w_t^i E_{ti}(y_t - B_i x_t - \mu_i)' \Sigma_i^{-1}(y_t - B_i x_t - \mu_i))$.
By a standard calculation and using equation 5, we show that the Implicit estimator of B_i is

$$\widehat{B_i}^{Imp} = (S_{YX',i} - \mu_i S_{X',i})(S_{XX',i})^{-1}. \qquad (14)$$

4.1.2. Tied parameters

For the tied case, we get

$$\widehat{B}^{Imp} = \sum_i (S_{YX',i} - \mu_i S_{X',i})(\sum_i S_{XX',i})^{-1}. \qquad (15)$$

4.2. Mean estimation

4.2.1. Untied parameters

The Implicit estimator of μ_i is a Normal distribution $\widetilde{p}(\mu_i|y_t, x_t, \Sigma_i, B_i) \propto \exp(-\frac{1}{2}\sum_t w_t^i E_{ti}(y_t - B_i x_t - \mu_i)' \Sigma_i^{-1}(y_t - B_i x_t - \mu_i))$.
Then, by using equation 5, we have

$$\widehat{\mu_i}^{Imp} = \frac{S_{Y,i} - B_i S_{X,i}}{\sum_t w_t^i}. \qquad (16)$$

If $B_i = 0$ (No regression), the estimator of μ_i becomes

$$\widehat{\mu_i}^{Imp} = \frac{S_{Y,i}}{\sum_t w_t^i}. \qquad (17)$$

4.2.2. Tied parameters

For the tied case, we get

$$\widehat{\mu}^{Imp} = \frac{\sum_i (S_{Y,i} - B_i S_{X,i})}{N}. \qquad (18)$$

4.3. Estimating the regression matrix and the mean simultaneously

Since both the equation for B_i and μ are mutually dependent on each other, we present the expression estimating them jointly. By applying the same method used in section 3.4, we get

$$\widehat{D_i}^{Imp} = S_{YZ',i} S_{ZZ',i}^{-1}. \qquad (19)$$

4.4. *Full Covariance matrix estimation*

4.4.1. *Untied parameters*

$\widetilde{p}(\Sigma_i|y_t,x_t,B_i,\mu_i) \propto |\Sigma_i|^{-\frac{w_i}{2}} \exp(-\frac{1}{2}\sum_t w_t^i E_{ti}(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)) \propto |\Sigma_i|^{-\frac{w_i}{2}} \exp(-\frac{1}{2}tr(\Sigma_i^{-1}\sum_t w_t^i E_{ti}(y_t - B_i x_t - \mu_i)(y_t - B_i x_t - \mu_i)')$

where $\widetilde{p}(\Sigma_i|y_t,x_t,B_i,\mu_i)$ is an Inverse-Wishart distribution with $(w_i - d - 1)$ degrees of freedom and a positive definite precision matrix
$\sum_t w_t^i E_{ti}(y_t - B_i x_t - \mu_i)(y_t - B_i x_t - \mu_i)'$.
Then, the Implicit estimator of Σ_i is given by

$$\widehat{\Sigma}_i^{Imp} = \frac{1}{w_i - 2d - 2}(A_i). \tag{20}$$

If there is no regression, the estimator of Σ_i becomes

$$\widehat{\Sigma}_i^{Imp} = \frac{S_{YY'} - S_{Y\mu_i'} - \mu_i S_{Y'} + \mu_i\mu_i'}{w_i - 2d - 2}. \tag{21}$$

4.4.2. *Tied parameters*

For the tied case, we have

$$\widehat{\Sigma}^{Imp} = \frac{1}{N - 2dN - 2N}(\sum_i A_i). \tag{22}$$

4.5. *Spherical covariance matrix estimation*

4.5.1. *Untied parameters*

If we have the constraint that $\Sigma_i = \sigma_i^2 I$ is isotropic, the conditional density of Y becomes
$p(y|x,Q=i) = c\sigma_i^{-d} exp(-\frac{1}{2}\sigma_i^{-2}\|y - B_i x - \mu_i\|^2)l = -d\sum_t\sum_i w_t^i \log|\sigma_i| - \frac{1}{2}\sigma_i^{-2}\sum_t\sum_i w_t^i E_{ti}\|y_t - B_i x_t - \mu_i\|^2$.
Hence,
$\widetilde{p}(y|x,Q = i) = \exp\Big(- d\sum_t\sum_i w_t^i \log|\sigma_i| - \frac{1}{2}\sigma_i^{-2}\sum_t\sum_i w_t^i E_{ti}\|y_t - B_i x_t - \mu_i\|^2\Big)$
where σ_i^2 follows an Inverse-Gamma distribution with a shape parameter $(\frac{w_i d}{2} - 1)$ and a scale parameter $(\frac{1}{2}\sum_t w_t^i E_{ti}\|y_t - B_i x_t - \mu_i\|^2)$.
We can easily deduce that the Implicit estimator of σ_i^2 is
$(\widehat{\sigma}_i{}^2)^{Imp} = \frac{\sum_t w_t^i E_{ti}(\|y_t - B_i x_t - \mu_i\|^2)}{dw_i - 4}$.
In order to compute the expected value of this distance, we use the fact that $x'Ay = tr(x'Ay) = tr(Ayx')$. So, $E[x'Ay] = tr(AE[yx'])$.
Hence,

$$(\widehat{\sigma}_i{}^2)^{Imp} = \frac{1}{dw_i - 4}tr(C_i). \tag{23}$$

4.5.2. *Tied parameters*

For the tied case, we have

$$(\widehat{\sigma}^2)^{Imp} = \frac{1}{Nd - 4N}tr(\sum_i C_i). \tag{24}$$

5. Comparative study

Table 2 provides us a summary of the estimators of conditional Gaussian distribution parameters (μ_i, B_i, Σ_i) obtained by Maximum of Likelihood and by Expectation A Posteriori (described in section 3) and our Implicit method (described in section 4).

First, we notice that there is a difference between the Implicit estimator and the classical ML one when estimating the covariance matrix, whereas both estimators coincide for the mean and regression parameters. According to our knowledge, there is no theoretical work which explains these coincidences. Concerning the comparison between Implicit and Bayesian estimation, we point out that the parameters estimated by Bayesian approach depend on the prior parameters and, as we know, the choice of a specific prior information has always been problematic, hence representing the major weakness of this approach. In most cases, we either need an expert to get the prior knowledge or, we have to use non informative priors.

Since Implicit and ML approaches coincide for the estimation of the mean parameter μ and also the parameter of regression B, we only have to compare them with the Bayesian (EAP) estimation of

	IMP	ML	EAP
μ_i	$\frac{S_{Y,i}-B_iS_{X,i}}{\sum_t w_t^i}$	$\frac{S_{Y,i}-B_iS_{X,i}}{\sum_t w_t^i}$	$(\psi^{-1}m+\Sigma_i^{-1}(S_{Y,i}-B_iS_{X,i}))(w_i\Sigma_i^{-1}+\psi^{-1})^{-1}$
B_i	$(S_{YX',i}-\mu_iS_{X',i})(S_{XX',i})^{-1}$	$(S_{YX',i}-\mu_iS_{X',i})(S_{XX',i})^{-1}$	$(\Sigma_i^{-1}(S_{YX',i}-\mu_iS_{X',i})+aV^{-1})\ (\Sigma_i^{-1}S_{XX',i}+V^{-1})^{-1}$
Σ_i	$\frac{A_i}{w_i-2d-2}$	$\frac{S_{YY',i}-S_{YZ',i}D_i'-D_iS_{ZY',i}+D_iS_{ZZ',i}D_i'}{w_i}$	$\frac{A_i+V}{w_i+\alpha-d-1}$

Table 2: Estimation of conditional Gaussian distribution parameters (μ_i, B_i, Σ_i) obtained by Implicit method (section 4), Maximum of Likelihood ([9]) and Expectation A Posteriori (section 3).

the covariance matrix. We can point out that the Implicit estimator of $\widehat{\Sigma}$ corresponds to the Bayesian one (EAP) by taking $V = 0$ and $\alpha = d-1$. However, this situation is impossible because V is a positive definite precision matrix. This situation is similar to the one described in section 2.2 for σ^2 estimation.

Since the parameter space is infinite, the Implicit distribution does not coincide with neither Fiducial nor Bayesian distribution. Therefore, our Implicit estimation should give more robust results than the Bayesian ones, particularly if the priors used in Bayesian estimation are far away from the true ones.

6. Experimentations

6.1. Experimental protocol

In order to evaluate the interest of using the Implicit approach for learning parameters in Conditional Gaussian Bayesian Networks and to measure the quality estimation, we have carried out repetitive experiments in several contexts.

In these contexts, we are able to control several parameters such as the number of variables n (n = 10, 30, 50) and the size of generated datasets N (N = 100, 1.000, 10.000). The maximal cardinality K of our discrete variables is also controlled for Conditional Gaussian Bayesian Networks (K = 2, 3, 5). In such conditions, every dataset generation is iterated 10x10 times, with 10 randomly generated DAGs, and 10 random parameter values for each of these DAGs.

Our goal is to compare the performance of two estimators working without any prior definition, i.e. the implicit and maximum likelihood estimators.

Our various models and algorithms have been implemented in Matlab with BNT [17]and BNT Structure Learning Package [18].

6.2. Evaluation criteria

Accuracy evaluation of each method is estimated by the Kullback-Leibler (KL) divergence between the "original" distribution used for generating a given dataset and the "final" distribution obtained with parameter learning. For large numbers of variable configurations (greater than 10^5), a Markov Chain Monte Carlo (MCMC) approximation is used with 10^5 random configurations.

Comparison of both methods is illustrated by plotting absolute values of KL obtained by the Implicit approach versus maximum likelihood for the same datasets. The fact that one method is better than the other can be observed with respect to the first diagonal (upper triangle : ML is better, versus lower triangle : implicit approach is better). In order to determine whether the observed differences are statistically significant, we use the Wilcoxon paired signed rank test, with a significance level equal to 0.05.

6.3. Results and interpretations

Figure 2 proposes the KL divergence obtained by Implicit approach versus the Maximum Likelihood one for the same datasets, for $n = 10$. Similar results have been obtained for $n = 30$ and 50.

Whatever the values of K (maximum cardinality of variables) and N (dataset size), the Implicit approach gives either similar or better results than the ML one. Both approaches coincide when N is high ($N = 1000$ and 10000, results in magenta and black) but also with a small sample size ($N = 100$) but only when the maximum cardinality is low ($K = 2$).

When the maximum cardinality is high ($K = 3$ and $K = 5$) and the dataset size N is low, the Implicit

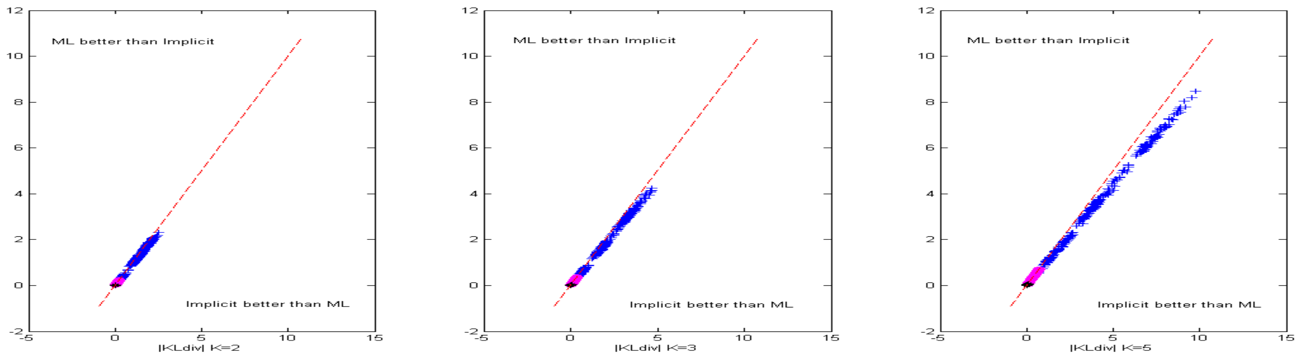

Figure 2: Comparison of KL divergence obtained by Implicit approach versus the method of maximum likelihood for the same datasets (upper triangle : ML is better, versus lower triangle : Implicit approach is better) with respect to dataset size ($N = 100, 1.000, 10.000$, resp. blue, magenta and black points) and maximum cardinality ($K = 2, 3, 5$).

approach gives more interesting results. All these results are also confirmed by the Wilcoxon tests which are not detailed here.

7. Conclusion and perspectives

In this paper, we introduce the notion of Implicit approach for the estimation of parameters in Conditional Gaussian Bayesian Networks (CGBNs). This method of estimation is similar to the Bayesian one, but happens in a natural context without specifying any prior parameters. This characteristic can be interesting for CGBNs where priors are not easily understandable or interpretable for users. Bayesian estimation with priors far away from the true values can lead to poor results. the Implicit (prior free) estimators proposed here are then very attractive to avoid such situations and to replace advantageously the ML estimator when the sample size is low.

This Implicit approach can also be used to learn the network structure. Most structure learning approaches use a score function that measures the goodness of fit between the structure and the data and thus try to find a good model optimizing this score. Many scoring functions have been proposed and are based on different principles, such as entropy or Bayesian approaches. Within this framework, our future work will propose an extension of Implicit score function proposed in [8] (and devoted to discrete BNs) in which CGBN structure inference can be based without determining any prior.

References

1. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
2. F.V. Jensen. *An introduction to Bayesian Networks.* Taylor and Francis, London, United Kingdom, 1996.
3. R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.
4. D. Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *J. Mach. Learn. Res.*, 5:1287 – 1330, 2004.
5. A. Hassairi, A. Masmoudi, and C Kokonendji. Implicit distributions and estimation. *Communications in Statistics - Theory and Methods*, 34(2):245 – 252, 2005.
6. H. Ben Hassen, A. Masmoudi, and A. Rebai. Causal inference in biomolecular pathways using a bayesian network approach and an implicit method. *Journal of Theoretical Biology*, 253(4):717 – 724, 2008.
7. H. Ben Hassen, A. Masmoudi, and A. Rebai. Inference in signal transduction pathways using em algorithm and an implicit algorithm: Incomplete data case. *Journal of Computational Biology*, 16(9):1227–1240, 2009.
8. H. Ben Hassen, L. Bouchaala, A. Masmoudi, and A. Rebai. Learning structure and parameters in bayesian networks using an implicit framework. In Ahmed Rebai, editor, *Bayesian Network*, pages 1–12. InTech, 2010.
9. K. P. Murphy. Fitting a conditional linear gaussian dis-

tribution. Technical report, UC Berkeley, Computer Science Division, 2003.

10. R.A. Fisher. The fiducial argument in statistical inference. *Ann. Eugen*, 6:391 – 398, 1935.
11. H.D. Heike, C. Târcolea, M. Demetrescu, and A.I. Tarcolea. Fiducial inference for discrete and continuous distributions. *Proceeding of the 2nd International Colloquium 'Mathematics in Engineering and Numerical Physics' Balan, V. ed.*, pages 69 – 80, 2003.
12. N. Mukhopadhyay. Bayesian inference some comments on Hassairi et al.s implicit distributions and estimation commun. *Statist.Theor. Meth*, 35:293 – 297, 2006.
13. S. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17:31–57, 1989.
14. K.G. Olesen. Causal probabilistic networks with both discrete and continuous variables. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:275–279, 1993.
15. D. Edwards. Hierarchical interaction models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 52:3–20, 1990.
16. D. Heckerman and D. Geiger. Learning bayesian networks: A unification for discrete and gaussian domains. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 274–284, San Francisco, CA, 1995. Morgan Kaufmann Publishers.
17. Kevin Murphy. The bayesnet toolbox for matlab. In *Computing Science and Statistics: Proceedings of Interface*, volume 33, 2001.
18. P. Leray and O. Francois. BNT structure learning package: Documentation and experiments. Technical report, Laboratoire PSI, 2004.