Journal de la Société Française de Statistique
Vol. 155 No. 3 (2014)

Crossed Linear Gaussian Bayesian Networks, parsimonious models

Titre: Réseaux Bayésiens Gaussiens Linéaires Croisés, des modèles parcimonieux

Simiao Tian[1], Marco Scutari[2] and Jean-Baptiste Denis[1]

Abstract: Linear Gaussian Bayesian networks can dramatically reduce the parametric dimension of the covariance matrices in the framework of multivariate multiple regression models. This idea is developed using structured, crossed directed acyclic graphs (DAGs) when node sets can be interpreted as the cartesian product of two sets. Some interesting properties of these DAGs are shown as well as the probability distributions of the associated Bayesian networks. A numerical experiment on simulated data was performed to check that the idea could be applied in practice. This modelling is applied to the prediction of body composition from easily measurable covariates and compared with the results of a saturated regression prediction.

Résumé : Dans cet article, nous proposons de substituer aux régressions linéaires multivariées classiques des sous modélisations plus parcimonieuses construites à l'aide de réseaux bayésiens gaussiens. L'idée est d'améliorer la prédiction de variables par des covariables, grâce à une réduction sensible de la dimension paramétrique de la matrice de variance-covariance. Une mise en œuvre est développée par l'utilisation de DAG (graphe orienté sans circuit) structurés lorsque l'ensemble des nœuds à modéliser est un produit cartésien de deux ensembles. Un certain nombre de propriétés intéressantes de ces DAG et des réseaux bayésiens associés en découle. Une expérimentation numérique basée sur des données simulées est réalisée pour vérifier la faisabilité de la proposition à partir de données lorsque la structure du DAG n'est pas connue. Enfin, la proposition est appliquée à la prédiction de la composition corporelle à partir de covariables faciles à obtenir. Les résultats obtenus par une recherche systématique de cette classe de réseaux bayésiens sont comparés avec la prédiction du modèle saturé de regression multiple multivariée.

Keywords: Bayesian network, crossed DAG, multivariate multiple regression, prediction
Mots-clés : réseau bayésien, DAG croisé, régression multiple multivariée, prédiction
AMS 2000 subject classifications: 62J05, 62P10, 62M20

1. Introduction

1.1. Starting from Multiple Regression

A very standard statistical model is the multivariate multiple regression model (Anderson, 2003, Chapter 8). Let us suppose that we have n observations with p variables to predict with the help of q covariates. The model reads

$$\begin{array}{rcl} \mathbf{Y} & = & \mathbf{X\Theta} + \mathbf{E} \\ V\left(vec\left(\mathbf{E}\right)\right) & = & \mathbf{I}_n \otimes \Sigma \end{array} \tag{1}$$

[1] Mathématiques et Informatique Appliquées, INRA, France.
E-mail: Jean-Baptiste.Denis@Jouy.Inra.Fr
[2] Genetics Institute, UCL, United Kingdom.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

where $\mathbf{Y}$ is the variable matrix $[n \times p]$, $\mathbf{X}$ is the covariate matrix augmented with a $\mathbf{1}$ vector $[n \times (1+q)]$, Θ is the expectation parameter matrix $[(1+q) \times p]$, $\mathbf{E}$ is the error matrix $[n \times p]$ and Σ is the covariance matrix $[p \times p]$. The number of parameters is $p(1+q)$ for the expectation and $p(p+1)/2$ for the covariance matrix. When p and q are large, n must be large as well to obtain estimates with desirable statistical properties. Of course, variances and covariances are more demanding in terms of sample size.

Many proposals have been made in the literature to offer more sophisticated and convenient statistical tools for multivariate regression problems. Some examples are the undirected graphical models used in Whittaker (1990), the multivariate analysis of variance (MANOVA) and seemingly unrelated regression (SUR) models in Timm (2002), the multivariate generalised linear models in Fahrmeir and Tutz (1994), and more recently the graphical lasso in Friedman et al. (2007).

The idea we develop in this paper is to use linear models in a more parsimonious framework based on linear Gaussian Bayesian networks (GBNs). Moreover when the structure of the set of the variables is crossed to use what we call crossed GBNs to decrease even more the number of parameters.

1.2. Linear Gaussian Bayesian Networks

1.2.1. Definition

Bayesian networks (BNs) are a class of probabilistic models used more and more in many fields of applications; at first they were developped for discrete variables but can be applied to any type of random variables. General presentations can be found in many books, for instance Naïm et al. (2004), Koller and Friedman (2009), Nagarajan et al. (2013) and Scutari and Denis (2014). A GBN is a BN (Neapolitan, 2003, sections 4.1.3 and 7.2.3; Korb and Nicholson, 2011, section 8.2) where every variable (or node) follows a Normal distribution. For each node, conditionally to its ascendants, the variance is constant and the expectation depends only on the direct parents through an affine transformation of the parent values. As a consequence, the joint probability distribution of the set of variables is multinormal with a free expectation and a constrained covariance matrix. In addition the acyclicity constraint of BNs induces a partial ordering on the nodes, and their relationships can be represented with a directed acyclic graph (DAG) (Pearl, 1988; Pearl, 2009; Leray, 2006, chapter 1; Koller and Friedman, 2009). More precisely, it exists at least one topological order on the node set, say $([1],[2],...,[p])$ such that the distributions can be defined by the following p equations:

$$Y_{[i]} \mid Y_{[1]},...,Y_{[i-1]} \sim N\left(\mu_{[i]} + \sum_{u=1}^{i-1} \rho_{[u],[i]} Y_{[u]}, \sigma^2_{[i]}\right) \text{ for } i = 1,...,p \tag{2}$$

where the summation term vanishes if node $Y_{[i]}$ has no parent. When the $p(p-1)/2$ regression coefficients $\rho_{[u],[i]}$ are all unknown and unconstrained, the GBN is saturated and there is no restriction on the form of the covariance matrix of the implied multinormal distribution. In that case, the model has $p(p+3)/2$ free parameters. If we denote the number of parents of the ith node with $p(i)$, there are p free parameters for the μs, p for the σs and $m = \sum_{i=1}^{p} p(i)$ for the ρs. It is easy to see that the μs and σs are respectively associated with the location and scale parameters

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

FIGURE 1. *A serial DAG with three nodes supporting the Bayesian network defined in (3).*

of the variables, so we can assume that all variables have marginal zero expectations and unity variances without altering the intrinsic properties of the model. As a result, the maximum number of parameters is $m = p(p-1)/2$, corresponding to the conveniently modified ρs and related to the $p(p-1)/2$ correlation parameters of the multinormal distributions.

1.2.2. Example

Just to give a small example, let us consider a GBN, based on the DAG drawn in Figure 1, with three marginally centred and normalised nodes with the following local distributions:

$$\begin{aligned} Y_1 &\sim N(0,1), \\ Y_2 \mid Y_1 &\sim N\left(\rho_{12}Y_1, \left(1-\rho_{12}^2\right)\right), \\ Y_3 \mid Y_1, Y_2 \sim Y_3 \mid Y_2 &\sim N\left(\rho_{23}Y_2, \left(1-\rho_{23}^2\right)\right); \end{aligned} \tag{3}$$

which imply the following joint distribution:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{12}\rho_{23} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}\right).$$

Compared to an unconstrained distribution on Y_1, Y_2 and Y_3, there is one less free parameter (ρ_{13}), due to the following constraint on the correlation matrix:

$$Cor(Y_1,Y_3) = Cor(Y_1,Y_2).Cor(Y_2,Y_3).$$

For any GBN, the number of free parameters in the correlation matrix is simply given by the number of arcs in the associated DAG, which is equal to m. It is important to note that this way to impose constraints on the correlation matrix is quite efficient and intuitive. However, expressing the induced constraints is not always as straightforward as in this small example, even though the rules to get the correlation coefficients from the regression coefficients can be expressed in a matricial closed form.

1.2.3. Matrix Formulation

To define the DAG associated with a GBN, it is convenient to use a $p \times p$ adjacency matrix (Nagarajan et al., 2013; Jungnickel, 2013), say A. Each row and each column of A is associated with one of the nodes in the DAG, and when Y_i is a parent of $Y_{i'}$, then $A_{i,i'}$ is equal to one, and zero

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

otherwise. Since it is equivalent to the DAG, the adjacency matrix shares all its properties; for instance, the number of arcs in the DAG is given by the sum of all the elements of A. Another more interesting property is that A^u provides the number of paths of length u joining any ordered pair of nodes built with successive arcs of the DAG (Bang-Jensen and Gutin, 2009). It is convenient to assume that the order of the rows and columns of the A matrix is a topological order of the nodes, that is one of the orders compatible with all arcs of the associated DAG; this implies that the A matrix is upper triangular with a null diagonal.

Any GBN can be defined with (i) the vector of the constants, say μ, (ii) the vector of the standard deviations, say σ and (iii) the matrix of the regression coefficients, say R, which has the same dimension and the same zeros as the adjacency matrix, but the regression coefficients instead of ones. Here are those matrices for Model (3):

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}; \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \sigma = \begin{pmatrix} 1 \\ \sqrt{1-\rho_{12}^2} \\ \sqrt{1-\rho_{23}^2} \end{pmatrix}; R = \begin{pmatrix} 0 & \rho_{12} & 0 \\ 0 & 0 & \rho_{23} \\ 0 & 0 & 0 \end{pmatrix}.$$

1.2.4. Joint Distribution

A basic problem is the computation of the joint distribution of the set of nodes, say Y, from the definition of the BN. Indeed, the definition of a GBN, with the marginal distributions of the root nodes and the conditional distributions for the others, provides only the local behavior of every node. For interpretative purposes, it is important to know the marginal distribution of non-root nodes, and the strength of the dependencies between indirectly related nodes. This is achieved with the joint distribution. In the multinormal framework, the point is to obtain the expectation vector and covariance matrix which is not always an easy task even for such tractable distributions. Two ways are reported in the following. The first relies on the topological order and is related to the algorithm illustrated in Korb and Nicholson (2011) section 2.4.1 for discrete BNs; also of interest are the proposals made by Neapolitan (2003) (section 4.1.3).

1. Affine transformation of white noise (a vector of independent centred and normalised normal variables), denoted by E, that is the identification of the vector M and matrix G such that $Y = M + G \cdot E$. This is obtained by (i) express the first node as $M_1 + G_{1,1}E_1$ and (ii) from the second node until the last node express $M_i + \sum_{u=1}^{i} G_{i,u}E_u$. Note that the matrix G is lower triangular and that all its diagonal components can be imposed to be strictly positive.

2. Use of the matrix R defined above by computing the matrix $R^* = \mathbf{I}_p + \sum_{u=1}^{p-1} R^u$. There are algorithms to compute it for a specific DAG (Bang-Jensen and Gutin, 2009; Sedgewick, 2011) which can be of interest when the number of nodes is large. Then, it can be checked that

$$E[\mathbf{Y}] = R^{*'} \cdot \mu \quad \text{and} \quad V[\mathbf{Y}] = R^{*'} \cdot \text{diag}(\sigma)^2 \cdot R^*. \tag{4}$$

   where $\text{diag}(\sigma)$ is the diagonal matrix built with vector σ.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

2. Crossed Gaussian Bayesian Networks

2.1. Definition

In some situations, the set of variables has a crossed structure, that is the variables can be indexed by a couple of indexes, all couples being present. In the following we will denominate these indexes: series of items. The most widely-known case is dynamic BNs (Ghahramani, 1997; Friedman et al., 1998), in which the same set of variables is observed at different successive times, but other situations are possible as shown in the examples below (§2.2). In order to obtain a parsimonious model, requiring only a small number of parameters, it can be profitable to use a crossed structure. To do so, we propose to use crossed DAGs, which are the product of one DAG on the first series of items by another DAG on the second series of items. In fact a crossed DAG is the Cartesian product of DAGs associated with each series of items (Bang-Jensen and Gutin, 2009). More formally, let us denote each variable with a pair of indices associated with the two series of items: $Y_{(i_1,i_2)}$ with $i_1 = 1,...,p_1$, $i_2 = 1,...,p_2$ and $p = p_1 p_2$; also be $A^{(1)}$ ($A^{(2)}$) the adjacency matrices associated onto the p_1 (p_2) items. The adjacency matrix of the resulting crossed DAG is given by the simple rule:

$$\begin{aligned} A_{(i_1,i_2),(j_1,j_2)} &= 1 \text{ when } \begin{cases} i_1 = j_1 & \text{and} \quad A^{(2)}_{i_2,j_2} = 1 \\ \quad \text{or} & \\ i_2 = j_2 & \text{and} \quad A^{(1)}_{i_1,j_1} = 1 \end{cases} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{5}$$

That is, for each set of nodes having a common item of series 1, the DAG for series 2 is applied; and conversely for each set of nodes having a common item of series 2, the DAG for series 1 is applied. Equation (5) is equivalent to the matrix formula

$$A = \mathbf{I}_{p_1} \otimes A^{(2)} + A^{(1)} \otimes \mathbf{I}_{p_2}.$$

From this formula, one can see that the number of ρ coefficients for a crossed BN is $p_1 m_2 + p_2 m_1$ where m_1 and m_2 are the number of ρs of the two elementary DAGs.

2.2. Examples

Crossing the DAG defined by (3) and shown in Figure 1 with itself produces the crossed DAG of Figure 2. This DAG can be used to propose a GBN, and the number of parameters is reduced from a maximum of 36 to 12. This situation occurs each time a multivariate observation is repeated in different correlated places. A reduced example can be that of different properties measured to the two eyes of an individual; the application on body composition, developped in §4, belongs to this category.

Now the DAG drawn in Figure 2 could as well be representing a dynamic BN with three slices of time (from one column to another) and a direct link for each variable of the system (associated with the rows). Notice that if dynamic BN nodes are described as the Cartesian product of two sets, crossed BNs apply only when temporal arcs are placed to each variable.

Other fields of application are two crossed random factor effects when the levels are not exchangeable: for instance a set of more or less related genotypes cultivated within different types of environments structured according to the year and soil categories.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

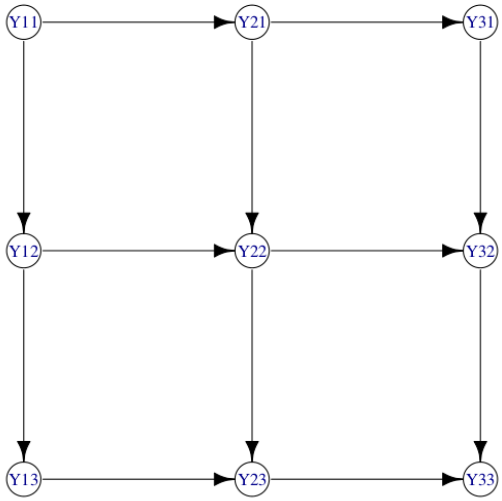

FIGURE 2. *Crossed DAG obtained by crossing the serial DAG shown in Figure (1) by itself.*

TABLE 1. *Different restrictions on the regression coefficients of a crossed DAG. p_1 and p_2 are the node numbers of the elementary DAGs generating the crossed DAG, and m_1 and m_2 are their respective parametric dimensions.*

type	constraints	parametric dimension
F.F	no one	$p_2m_1 + p_1m_2$
C.F	identical for each item of series 1	$p_2m_1 + m_2$
F.C	identical for each item of series 2	$m_1 + p_1m_2$
C.C	identical for both series	$m_1 + m_2$

2.3. Additional Constraints

In order to decrease even more the parametric dimension, some constraints linked to the crossed structure can be added on the regression coefficients. Particularly, some equalities can be imposed, like those implied by:

$$R = \mathbf{I}_{p_1} \otimes R^{(2)} + R^{(1)} \otimes \mathbf{I}_{p_2}$$

where $R^{(1)}$ ($R^{(2)}$) is some regression matrix associated with the DAG of the first (second) series of items. In that case, the number of ρ coefficients is just $m_1 + m_2$, which can be a drastic drop. Intermediate proposals can be made, examples are given in Table 1.

2.4. Introducing Covariates

2.4.1. Introduction

When discussing GBNs in the previous sections, we focused only on the variables, Y. We will now incorporate covariates, X, to match the regression model described in Equation (1). The basic idea is to add the covariates as *ancestors* of the variables in the crossed DAG and then retain the sub BN conditionally to the covariates. Starting from the simple example in Figure 2, an example

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

is shown in Figure 3-ii. The presence of the conditioning covariates alters the properties of GBNs previously indicated. For instance, the expectations cannot be further supposed to be null since they depend on the covariates' values; in addition, the covariance matrix loses the simplicity of Equation (4).

2.4.2. *Example*

As an example of the increased complexity introduced by the covariates, consider a toy example of one covariate intervening in two nodes of a 2×2 crossed DAG. Suppose that the joint distribution between variables and covariates can be described with a centred and normalised GBN as proposed in Figure 3-i, that is:

$$\begin{array}{rcc} C & \sim & N(0,1) \\ Y_{1,1} \mid C & \sim & N\left(eC, 1-e^2\right) \\ Y_{1,2} \mid Y_{1,1} & \sim & N\left(aY_{1,1}, 1-a^2\right) \\ Y_{2,1} \mid Y_{1,1} & \sim & N\left(cY_{1,1}, 1-c^2\right) \\ Y_{2,2} \mid C, Y_{1,2}, Y_{2,1} & \sim & N\left(fC+dY_{1,2}+bY_{2,1}, \sigma^2_{2,2}\right) \end{array}$$

where

$$\sigma^2_{2,2} = \quad 1-\left(f^2+d^2+b^2+2(efad+efcb+adcb)\right).$$

In the equations above, the main difficulty lies in defining the conditional variances to achieve all the marginal variances to be one. The structure of the covariance (here correlation) matrix is more evident, giving only the upper part:

$$V\begin{bmatrix} C \\ Y_{1,1} \\ Y_{2,1} \\ Y_{1,2} \\ Y_{2,2} \end{bmatrix} = \begin{pmatrix} 1 & e & ce & ae & f+ade+bce \\ - & 1 & c & a & ef+ad+bc \\ - & - & 1 & ac & cef+acd+b \\ - & - & - & 1 & aef+d+abc \\ - & - & - & - & 1 \end{pmatrix}.$$

It can be checked that every correlation is the sum of the products of the regression coefficients following every possible path linking the two considered nodes. This is a consequence of Equation 4.

The induced regression formulas can be computed from the equations above as follows:

$$E\left[\begin{pmatrix} Y_{1,1} \\ Y_{2,1} \\ Y_{1,2} \\ Y_{2,2} \end{pmatrix} \mid C\right] = \begin{pmatrix} e \\ ce \\ ae \\ f+ade+bce \end{pmatrix} \cdot C,$$

$$V\left[\begin{pmatrix} Y_{1,1} \\ Y_{2,1} \\ Y_{1,2} \\ Y_{2,2} \end{pmatrix} \mid C\right] = \begin{pmatrix} 1-e^2 & c(1-e^2) & a(1-e^2) & (ad+bc)(1-e^2) \\ - & 1-c^2e^2 & ac(1-e^2) & acd(1-e^2)+b(1-c^2e^2) \\ - & - & 1-a^2e^2 & abc(1-e^2)+d(1-a^2e^2) \\ - & - & - & 1-(f+ade+bce)^2 \end{pmatrix}.$$

Some of these expressions can be interpreted in terms of paths over the DAG from Figure 3-i, but others are more complicated and have no obvious graphical interpretation. The presence of two

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

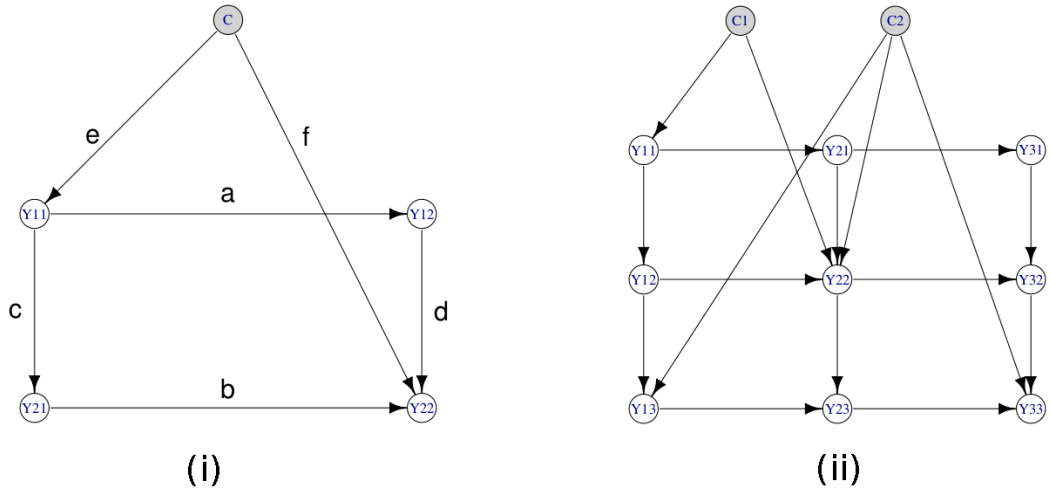

FIGURE 3. *(i) Toy* 2×2 *crossed DAG with one covariate (C); regression coefficients of the centred normalised distribution are indicated on each arc of the DAG. (ii) Crossed DAG from Figure 2 completed with two covariates (C*1 *and C*2*). The covariates intervene only on some of the variables for parsimony.*

or more covariates (Figure 3-ii) makes the algebraic computations intractable, only numerical computations are possible in general.

2.5. *Parametric Dimension*

Previous considerations showed that the parametric dimension of a GBN is $2p+m$ where p is the number of nodes and m the number of arcs. They are respectively associated with the constant terms of the expectations, the conditional variances and the regression coefficients. When considering a GBN comprising p variables and q covariates with the restriction of no variable being parent of some covariate, the parametric dimension is $2(q+p)+m_{XX}+m_{XY}+m_{YY}$ where m_{XX}, m_{XY} and m_{YY} are respectively the number of arcs within the variables, between a covariate and a variable, and within the covariates. Let us write down the joint density of this BN $[X,Y]$, we have that

$$\begin{aligned} [X,Y] &= [X]\,[Y \mid X] \\ pd\,([X,Y]) &= pd\,([X]) + pd\,([Y \mid X]) \\ 2\,(q+p)+m_{XX}+m_{XY}+m_{YY} &= (2q+m_{XX}) + pd\,([Y \mid X]) \end{aligned}$$

where $pd([Z])$ is the parametric dimension of the density $[Z]$. It follows that the parametric dimension of the conditional GBN is $2p+m_{XY}+m_{YY}$.

The decomposition of the joint distribution into $[X]$ and $[Y|X]$ has two interesting consequences. First when one is interested in the variables, there is no need to introduce arcs between covariate nodes when drawing the DAG. Second, the distributions of the covariates need not to be of the

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

same type that the distributions of the variables as we supposed here; even the nature of covariates could be different.

2.6. Parameter Estimation

When a GBN is free with respect to its DAG, that is the parameters of each Equation (2) are not constrained, maximum likelihood (ML) estimates can be computed with the standard ML estimators of each equation.

When some simple additional equalities are assumed like those in Table 1, things are more difficult. ML estimators cannot be simply obtained from data by stacking the variables and the corresponding parents since the same regression coefficient can be involved in different sets of regression coefficients. The situation is similar to that of dynamic GBNs, but without the additional complication of handling hidden variables (Murphy, 2002, Chapters 3 and 4). We did not find reference to estimate parameters in these circumstances for GBNs, only for discrete BNs (see for instance Section *17.5 Learning Models with Shared Parameters* in Koller and Friedman, 2009). To overcome the difficulty, we adapted the following heuristic alternating least squares (for instance described in Lütkepohl, 2005) procedure:

- Define as score for the difference between two GBNs sharing the same DAG, the sum of squared discrepancies of all the parameters, including the standard deviation.
- Initialise all the parameter values with the unconstrained fit.
- Iterate until the decrease of the score be less that some predefined threshold. Each iteration is a cycle over all expectation parameters. Each expectation along with the parameters and the standard deviations of the involved nodes is updated in turn, while keeping all the others fixed. Estimation is performed by weighted least squares since each elementary fit is associated with the fit of a linear model.

For all the examples we considered, convergence was very fast, typically after less than ten iterations. The implementation of this algorithm is included within the RBMN R package (Denis, 2013) available on CRAN but further research on estimation is required.

3. Numerical Experiment

In order to check that the models we proposed could be efficiently tackled, a small numerical experiment was performed. Indeed, using GBNs seems a bright idea to decrease the model complexity of multivariate multiple regressions but the challenge of discovering the true structure of the DAG could compromise the quality of the prediction and the remedy could be worst than the disease. Numerically checking the quality of such a general process is a huge task due to the numerous causes involved and we had to make some choices. After describing which simulations were performed, how predictions were made and how they have been assessed, the results are proposed in four tables and commented in a last subsection.

All calculations have been done with R (R Core Team, 2013) with the help of the three packages BNLEARN (Scutari, 2010), RBMN (Denis, 2013) and GLMNET (Friedman et al., 2010).

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

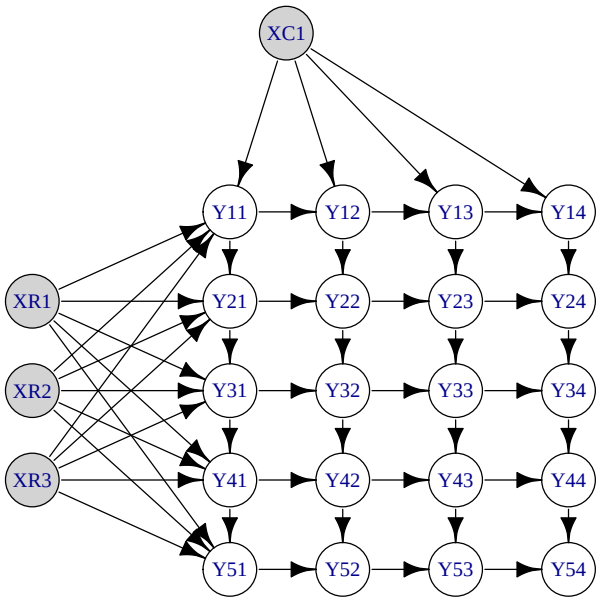

FIGURE 4. *The DAG of a model used for the numerical experiment when* $nrow = 5$, $ncol = 4$, $xrow = 3$ *and* $xcol = 1$.

3.1. Simulation Setup

3.1.1. Model Scenarios

GBNs provided by function `generate8grid7crossed` of the rbmn package were used; a typical associated DAG is shown in Figure 4. Variables follow a grid structure of *nrow* rows and *ncol* columns; arcs link successive variables within rows (and columns) from left to right (and top to bottom). There are two series of covariates: *xrow* row covariates and *xcol* column covariates; the first series is assigned to all variables of the first column, the second series is assigned to all variables of the first row. Then a model is mainly defined by the following parameters:

nrow,ncol the dimensions of the set of crossed variables, in the presented calculations we used $nrow = 10$ and $ncol = 5$,

xrow,xcol the numbers of independent covariates for the first column and first row variables,

ρ.xr the matrix values [xrow,nrow] for the regression coefficients, associated with the arcs between a row covariate and the first column variable,

ρ.xc the matrix values [xcol,ncol] for the regression coefficients, associated with the arcs between a column covariate and the first row variable,

ρ.yr the matrix values [nrow,ncol-1] for the regression coefficients, associated with the arcs between successive variables on the same row,

ρ.yc the matrix values [ncol,nrow-1] for the regression coefficients, associated with the arcs between successive variables on the same column.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

All local (marginal or conditional) distributions are given a null expectation and unit variance. To reduce the computation, all ρ coefficients were used with the same unique value.

3.1.2. Ten Different Strategies for Prediction

The aim is to check whether using a GBN whose structure and parameters are learned from a data set can perform better than a standard multivariate multiple regression for prediction. To be fairer to the standard regression, we implemented three estimations to build the predictor: the standard least squares, the lasso and the elastic net, respectively coded as S-ls, S-la and S-en with the help of the GLMNET package (Friedman et al., 2010). The known true model was also implemented being the target to tend to (denoted as G-true), in fact only the DAG was used, the parameters being estimated from the data sets. Finally, six GBN predictions were attempted, differing in the amount of information provided to the model structure; also with two different ways of indicating that some nodes of the DAG are covariates and cannot be children of the variables. They are:

1. Learning the DAG constraining (i) every covariate to be a parent of every variable and (ii) preventing arcs between covariates [G-c],
2. Like (1) and preventing $Y(i',j')$ to be a parent of $Y(i,j)$ when $j<j'$ [G-cc],
3. Like (2) and preventing $Y(i',j')$ to be a parent of $Y(i,j)$ when $(i<i')$ [G-ccr],
4. Learning the DAG constraining (i) covariates not to be children of any variable and (ii) preventing arcs between covariates [G-n],
5. Like (4) and preventing $Y(i',j')$ to be a parent of $Y(i,j)$ when $j<j'$ [G-nc],
6. Like (5) and preventing $Y(i',j')$ to be a parent of $Y(i,j)$ when $(i<i')$ [G-ncr].

All results presented below were obtained with the tabu search of the `tabu` function of BNLEARN R package (Scutari, 2010) using the default parameter values.

3.1.3. Simulation Design

Any simulation is a two-levels simulation of data sets of N observations comprising the variables and covariates, according to the two repetition numbers: S and T.

A data set is a triplet of related matrices $\left(X,Y^{TR},Y^{VA}\right)$ of respective sizes $(N\times C)$, $(N\times V)$ and $(N\times V)$ where N is the sample size, C the number of covariates and V the number of variables. Y^{TR} is used for the TRaining, Y^{VA} being reserved for the VAlidation. Such triplets of matrices are generated ST times and denoted $\left(X_s,Y_{st}^{TR},Y_{st}^{VA}\right)$ for $s=1,...,S$ and $t=1,...,T$ evidencing the fact that covariates, X_s, are generated only S times and are identical for the training and validating data sets, while $2T$ repetitions for each covariate configuration is done for the variables. This design is necessary (i) to get a fair assessment of the error of prediction and (ii) to distinguish the bias from the standard deviation.

For such a simulation, a predictor (of the variables using the covariates) is obtained independently from $\left(X_s,Y_{st}^{TR}\right)$ on each st simulation and applied for a prediction to the associated data set $\left(X_s,Y_{st}^{VA}\right)$. Let us denote it by P_{st}, it is a $(N\times V)$ matrix.

As the model is known, the true target values are known, It is the matrix of conditional expectations of size $(N\times V)$, different for each draw of the covariates. Let us denote it by Λ_s.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

To assess a prediction we will use the following ST matrices of the discrepancies between the prediction and the conditional expectation:

$$D_{st} = P_{st} - \Lambda_s.$$

3.1.4. *Criteria for comparison*

Bias Let $D_{s.}$ be the mean over t on matrices D_{st}, it can be interpreted as the estimated bias for the sth draw of the covariates. And we will use

$$\begin{aligned} B_v^2 &= \frac{\sum_{s=1}^{S}\sum_{n=1}^{N}\left((D_{s.})_{nv}\right)^2}{SN} \\ B^2 &= \frac{\sum_{v=1}^{V} B_v^2}{V} \end{aligned}$$

as squared bias for each variable and global squared bias.

Standard Deviation In the same way, the dispersion around the average predicted values will be assessed with

$$\begin{aligned} SD_v^2 &= \frac{\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{n=1}^{N}\left((D_{st} - D_{s.})_{nv}\right)^2}{STN} \\ SD^2 &= \frac{\sum_{v=1}^{V} S_v^2}{V} \end{aligned}$$

S.E.P. Finally, the standard error of prediction can be written as

$$\begin{aligned} SEP_v^2 &= \frac{\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{n=1}^{N}\left((D_{st})_{nv}\right)^2}{STN} \\ SEP^2 &= \frac{\sum_{v=1}^{V} SEP_v^2}{V} \end{aligned}$$

3.2. *Simulation Sizes*

These simulations are quite time consuming and a first study was conducted to determine adapted values for S and T. Table 2 presents the three criteria for $N = 100$, $xrow = 2$, $xcol = 1$ and $\rho = 0.5$. Even for the smallest combination ($S = 10, T = 20$) results are similar to the most expansive one ($S = 30, T = 80$). Therefore, for computational reasons, all further calculations were made with the intermediate combination ($S = 20, T = 30$).

Another important parameter is the sample size of individual data sets, that is N. The S.E.P. was computed for the ten predictions with different values varying from 50 to 500 (see Table 3). It appears that there is a strong effect of N within each prediction while the comparisons between predictions remain similar for a fixed sample size. Again for computational reasons, we retained the intermediate value of $N = 100$.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

TABLE 2. *Effect of the repetition numbers, S and T, on the calculations of the criteria. The other parameters are fixed to nrow* $= 10$, *ncol* $= 5$, *xrow* $= 2$, *xcol* $= 1$, $N = 100$ *and* $\rho = 0.5$.

		S-ls			G-true		
$T =$		20	30	80	20	30	80
$S = 10$	bias	0.069	0.054	0.032	0.032	0.028	0.018
	SD	0.289	0.293	0.295	0.144	0.148	0.148
	SEP	0.297	0.298	0.297	0.147	0.151	0.149
$S = 20$	bias	0.069	0.055	0.032	0.031	0.027	0.017
	SD	0.290	0.293	0.296	0.144	0.148	0.148
	SEP	0.298	0.298	0.297	0.147	0.151	0.149
$S = 40$	bias	0.065	0.055	0.032	0.031	0.027	0.017
	SD	0.291	0.294	0.296	0.145	0.148	0.148
	SEP	0.298	0.299	0.298	0.149	0.150	0.149

TABLE 3. *Looking for the sample size, N, effect on the S.E.P.. The other parameters are fixed to nrow* $= 10$, *ncol* $= 5$, *xrow* $= 2$, *xcol* $= 1$, $S = 20$, $T = 30$ *and* $\rho = 0.5$.

N	S-ls	S-la	S-en	G-c	G-n	G-cc	G-nc	G-ccr	G-ncr	G-true
50	0.422	0.420	0.420	0.338	0.412	0.319	0.390	0.295	0.364	0.213
60	0.384	0.382	0.382	0.308	0.362	0.283	0.338	0.264	0.312	0.195
70	0.359	0.357	0.357	0.283	0.337	0.256	0.303	0.239	0.280	0.179
80	0.333	0.332	0.332	0.261	0.311	0.236	0.278	0.220	0.253	0.168
90	0.314	0.312	0.312	0.249	0.295	0.220	0.254	0.297	0.232	0.162
100	0.298	0.297	0.297	0.234	0.272	0.204	0.234	0.192	0.214	0.151
200	0.210	0.209	0.209	0.171	0.183	0.132	0.140	0.127	0.132	0.105
300	0.172	0.171	0.171	0.144	0.141	0.106	0.107	0.102	0.103	0.086
400	0.149	0.148	0.148	0.132	0.122	0.089	0.090	0.087	0.087	0.074
500	0.134	0.133	0.133	0.121	0.108	0.080	0.080	0.078	0.077	0.068

3.3. *Effect of the Arc Strength*

Once the numbers of repetitions have been fixed, we first varied the dependence between directly related nodes changing the ρ value from 0 to 1: resulting S.E.P.s are presented in Table 4. Notice that when the regression coefficients are all zero, variables are independent from the covariates (also mutually independent) with null expectation and variance unity; this explains the null S.E.P. for the true model since it does not depend anymore on the pseudo-randomly generated conditioning covariates. It is amazing to see how the S.E.P. increases with ρ for all the predictions. This is the simple consequence of the increase of the marginal variances of nodes in the last rows and last columns due the multiplicative effect of the variation transmission as shown in the last line of Table 4.

3.4. *Effect of the Covariate Numbers*

The S.E.P values obtained for the ten predictions for different numbers of covariates have been collected in Table 5.

3.5. *Conclusions*

From the performed numerical experiments (see Tables 4 and 5) some conclusions appear:

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

TABLE 4. *ρ effect on the S.E.P. for the ten predictions (first 10 lines); marginal standard deviation of four variables according to the same variation of parameter ρ (last 4 lines). The other parameters are fixed to nrow = 10, ncol = 5, xrow = 2, xcol = 1, S = 20, T = 30 and N = 100.*

(SEP)	0	0.25	0.5	0.75	1
S-ls	0.200	0.212	0.298	2.119	40.589
S-la	0.194	0.208	0.297	2.118	40.592
S-en	0.194	0.209	0.297	2.117	40.592
G-c	0.115	0.131	0.234	2.278	35.552
G-n	0.090	0.157	0.272	2.228	39.351
G-c.cc	0.111	0.129	0.204	1.852	34.803
G-nc	0.086	0.160	0.234	2.255	36.596
G-c.ccr	0.110	0.125	0.192	1.685	33.674
G-ncr	0.083	0.158	0.214	1.804	34.744
G-true	0.000	0.080	0.151	1.562	33.093
St.Dev.(Y.1.2)	1	1.031	1.118	1.250	1.414
St.Dev.(Y.1.5)	1	1.033	1.154	1.469	2.236
St.Dev.(Y.3.5)	1	1.074	1.599	5.184	20.761
St.Dev.(Y.10.5)	1	1.075	1.856	36.861	1020.685

- As expected, the true DAG [G-true] always gives the best prediction.
- Sophisticated estimations (lasso and elastic net) are not decisively better than the standard least-squares estimations for the multivariate multiple regression model.
- Predicting with the considered GBN models always gives better prediction than using the saturated regression model.
- Among the six GBN models, there are differences which depend on the circumstances:
  - The strongest difference is between the two ways used to indicate the covariate nodes, but the difference depends on the strength of the relationships. When the strength is low it is worth imposing all covariates as parents of all variables, even if most of them does not exist. The explanation seems that when the strength is high, the learning algorithm may be able to recover the structure and not too many arcs are missed.
  - Adding information about the ordering of the columns and rows improved predictions but not as much as using a GBN instead of the saturated model. This is encouraging because it means that the strongest point is to use a GBN approach for predicting.
  - When the number of covariates increase in the model, surprisingly the prediction are worst. This could be the effect of different marginal variances due to more randomness? The reassuring point is that the main differences between the different predictions are not modified.

This is positive but it must be indicated that, from results not shown, not all learning algorithms provided good results. Other attempts with constraint-based algorithms like *grow-shrink* gave bad results, much worse than the saturated regression. Detailed examination of the resulting DAG showed that in a non negligible proportion of simulations, arcs were not discovered, thus introducing bias that was obviously absent from saturated models. Increasing the α parameter to retain more arcs was not the solution since it implied graphs not reducible to DAGs. The present positive results were obtained with the score-based algorithm more precisely the *tabu* algorithm implemented in the BNLEARN R package (Scutari, 2010) and used with the default parameters of its `tabu` function.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

TABLE 5. *Effect of the number of row/column covariates on the S.E.P. for the ten predictions. The other parameters are fixed to nrow* $= 10$*, ncol* $= 5$*,* $S = 20$*,* $T = 30$*,* $N = 100$ *and* $\rho = 0.5$*.*

xrow	xcol	S-ls	S-la	S-en	G-c	G-n	G-cc	G-nc	G-ccr	G-ncr	G-true
1	0	0.212	0.211	0.211	0.157	0.172	0.133	0.139	0.124	0.130	0.105
1	1	0.258	0.257	0.257	0.202	0.224	0.176	0.183	0.161	0.164	0.123
1	2	0.298	0.297	0.297	0.229	0.256	0.202	0.210	0.184	0.189	0.134
1	3	0.335	0.333	0.333	0.252	0.280	0.225	0.235	0.206	0.217	0.146
3	0	0.298	0.297	0.297	0.216	0.176	0.199	0.268	0.196	0.258	0.161
3	1	0.335	0.334	0.334	0.255	0.325	0.230	0.304	0.220	0.284	0.173
3	2	0.366	0.364	0.364	0.281	0.343	0.250	0.317	0.237	0.297	0.181
3	3	0.396	0.394	0.394	0.296	0.368	0.268	0.337	0.255	0.317	0.190
5	0	0.366	0.365	0.365	0.259	0.405	0.248	0.400	0.243	0.395	0.198
5	1	0.396	0.394	0.394	0.293	0.435	0.270	0.429	0.263	0.412	0.206
5	2	0.422	0.420	0.420	0.317	0.445	0.287	0.444	0.278	0.422	0.216
5	3	0.449	0.447	0.447	0.333	0.478	0.305	0.466	0.294	0.441	0.225
7	0	0.422	0.421	0.421	0.292	0.532	0.286	0.520	0.281	0.518	0.229
7	1	0.449	0.447	0.447	0.325	0.561	0.308	0.565	0.302	0.547	0.238
7	2	0.471	0.469	0.469	0.346	0.588	0.322	0.592	0.313	0.559	0.243
7	3	0.494	0.492	0.492	0.362	0.596	0.336	0.608	0.327	0.565	0.250

4. Application

4.1. Presentation

The human body composition is the allocation of body weight among three components: (L)ean, (F)at and (B)one. In detailed analyses, the body composition is investigated for each of the main segments of the body: (T)runk, (L)egs and (A)rms; so nine variables crossing the three components with the three segments are available. Body composition is an important diagnostic indicator since ratios of these masses can reveal regional physiological disorders. In the following, we will try to predict it from easily accessible covariates: the (A)ge in years, the (H)eight in cm, the (W)eight in kg and the waist (C)ircumference in cm; more details can be found in Tian et al. (2013) where a saturated model was used. For this purpose, we retained a data set of one hundred white men available in the RBMN R package (Denis, 2013). For each individual the variables to predict as well as the covariates have been recorded. Below are the first four individuals.

```
   A   H     W   C   TF   LF  AF   TL   LL   AL  TB  LB  AB
1 83 182  92.6 117 17.1  8.9 3.0 31.2 18.5  6.6 0.6 1.1 0.5
2 68 169  74.7  93  8.3  5.4 2.0 28.0 16.2  7.5 0.7 1.0 0.5
3 28 182 112.2 112 17.7 11.3 3.1 36.7 24.5 10.1 0.8 1.1 0.5
4 41 171  82.6  96 10.6  6.5 1.8 29.2 19.6  7.8 0.8 1.1 0.5
```

Here, (LF) stands for the (L)eg (F)at, and so on; all variables are given in kg. An additional covariate, the body mass index (B) has been calculated; it is a very popular score normalising the weight by the height. Overall we have $q = 5$ covariates and $p = 9 = 3 \times 3$ variables for $n = 100$ individuals.

All calculations have been done with R (R Core Team, 2013) with the help of the three packages BNLEARN (Scutari, 2010), RBMN (Denis, 2013) and GLMNET (Friedman et al., 2010). DAGs have been drawn using the IGRAPH (Csardi and Nepusz, 2006) package as all other DAGs of the paper.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

4.2. *Predicting with a crossed BN*

Using the crossed structure of the nine variables to perform the prediction, we tried to improve on that given by the standard multivariate multiple regression model from Equation (1). We are then in a model selection perspective (Claeskens and Hjort, 2008 and Burnham and Anderson, 2010). To do so, following Hastie et al. (2009), we randomly splitted our data set into two subsets of size $n = 50$, one for estimating models and the second one to assess their predictive power, the same splitting was used for all models. The validation of the models over an independent subsample exempted us from considering model selection criteria like AIC. In the framework of linear regression, the prediction of a new individual is made with a Normal distribution with expectation obtained from the regression formula and with standard deviation given by the estimation of the standard error. The difference between the observed value and the expectation is the bias (B_i^v, for the individual i and variable v); the squared standard deviation is the variance of the prediction ($(\sigma^v)^2$). To obtain a global score, we computed a global bias, a global standard deviation and a global standard error of prediction by summing them up as follows:

$$\begin{aligned} |Bias| &= \sum_{v=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} |B_i^v| \right), \\ Sd.Dev. &= \sum_{v=1}^{p} \sigma^v, \\ SEP &= \sum_{v=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{|B_i^v|^2 + (\sigma^v)^2} \right). \end{aligned} \tag{6}$$

In addition to these global quality scores, we measured the parametric dimensions with the number of arcs linking the covariates to the variables (m_{XY}: the number of retained regression coefficients) and the number of arcs between pairs of variables (m_{YY}: related to the complexity of the correlation structure within the variables). Also, we introduced $\tilde{m}_{YY}$, the parametric dimension of the covariance matrix, which is smaller in case of equality constraints on the regression parameters.

A systematic series of non degenerated crossed BNs were attempted among all possible 25 DAGs within the three compartments (F, L, B) crossed with all possible 25 DAGs within the three segments (T, L, A), and for each one the four constraint types proposed in Table 1. Table 6 shows the results for the best twelve, together with the corresponding results for the saturated model. Some interesting features can be noticed from this table:

- The selected crossed BNs perform better than the saturated model globally (*SEP*) and for both the bias and the variance, the improvement being greater for the variance.
- The reduction in the number of parameters with respect to the saturated model is striking, especially for the variance parameters.
- None of the selected models is without constraints (type F.F).
- For all the selected models, constraints on the three segments (Trunk, Legs, Arms) are present.
- For the segments, only two generating BNs (numbers 14 and 6) are present among the possible 25 ones. These two generating BNs are nested since 14 is (A $\rightarrow$ T $\rightarrow$ L) and 6 is

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

TABLE 6. *Quality prediction of the saturated model and the best 12 found crossed BNs. BN-c and BN-s are the coding of the generating BNs for the two series (compartment and segment). The constraint type refers to Table 1. |Bias|, Sd.Dev and SEP are defined in (6). m_{XY} is the number of arcs from a covariate to a variable; m_{YY} is the number of arcs from a variable to another variable; $\tilde{m}_{YY}$ is the parametric dimension from the m_{YY} arcs (the number of constraints have been subtracted).*

model	BN-c	BN-s	constraint type	$\lvert Bias\rvert$	*Sd.Dev*	*SEP*	m_{XY}	m_{YY}	$\tilde{m}_{YY}$
S-ls	-	-	-	5.97	7.51	10.19	45	36	36
S-la	-	-	-	5.95	7.51	10.17	45	36	36
S-en	-	-	-	5.95	7.51	10.17	45	36	36
1	22	14	F.C	5.82	6.98	9.66	21	15	9
2	12	14	F.C	5.82	6.99	9.67	19	12	8
3	3	14	F.C	5.82	7.00	9.68	21	9	7
4	9	14	F.C	5.83	7.01	9.69	20	12	8
5	12	6	F.C	5.80	7.05	9.70	20	9	5
6	22	6	F.C	5.80	7.05	9.70	22	12	6
7	12	14	C.C	5.81	7.05	9.72	19	12	4
8	22	14	C.C	5.82	7.06	9.73	21	15	5
9	3	6	F.C	5.82	7.07	9.73	23	6	4
10	9	6	F.C	5.83	7.07	9.74	22	9	5
11	11	14	F.C	5.88	7.03	9.74	19	12	8
12	12	6	C.C	5.81	7.09	9.74	20	9	3

(A; T → L).

– For the compartments, two generating BNs (numbers 22 and 12) are the most frequent. Also these two generating BNs are nested since 12 is (B ← L → F) and 22 adds to it the arc (F → B).

For the sake of the example, consider model 2 from Table 6, obtained with the combination of the preponderant generating BNs, the twelfth for the compartments and the fourteenth for the segments. Here are its regression equations (the conditional standard deviations are reported in parentheses):

```
AL = 5.133 + 0.139*W + -0.093*C  (0.687)
TL = -13.854 + 0.131*H + 0.156*W + 1.039*AL  (1.308)
LL = 9.418 + -0.026*A + 0.206*W + -0.125*C + 0.2*TL  (1.244)
AB = -0.359 + 0.004*H + 0.011*AL  (0.049)
TB = 0.219 + -0.003*C + 0.011*TL + 0.915*AB  (0.072)
LB = 0.271 + 0.011*LL + 0.835*TB  (0.097)
AF = 1.899 + 0.003*A + -0.025*H + 0.091*W + -0.387*AL  (0.409)
TF = 104.054 + -0.686*H + 0.918*W + 0.283*C + -2.456*B + 0.432*AF + -0.387*TL  (1.247)
LF = -3.71 + -0.018*A + 0.136*W + 0.248*B + 0.027*TF + -0.387*LL  (1.279)
```

The corresponding DAG is presented in Figure 5. Such a simple model displays good predictive power, and can also be used as a starting point for understanding the phenomenon under investigation. It makes sense for the (L)ean compartment be to the origin of most of the variation compared to the (F)at and (B)one compartments; and that given the (T)runk composition, there is no more correlation between the (A)rm and (L)eg segments.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

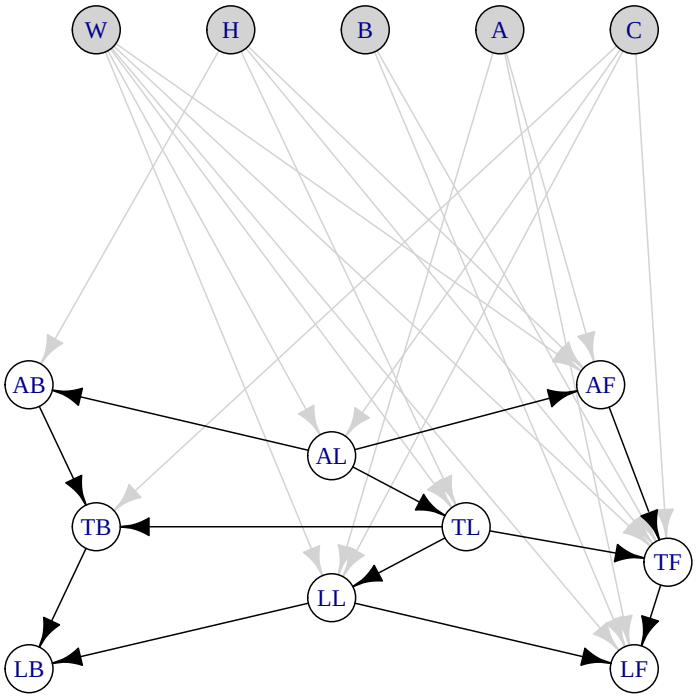

FIGURE 5. *DAG associated with Model 2 of Table 6. Arcs within covariates were not drawn for the sake of clarity; they are of no importance when conditioning by the covariates.*

5. Discussion

ANOVA models, regression models and their combinations presented in the framework of linear models are versatile tools for analysing complex data sets at the level of big trends, that is at the level of modelling the expectations of random variables (Graybill, 1976). To achieve better predictions, in this framework the common way is to reduce the number of used covariates looking for a small and efficient subset (see Miller, 2002 and Celeux et al., 2006 for a review). The next step is the modelling of variances and covariances. Random models are the natural extensions and many developments have been proposed in that direction, such as the introduction of variance components in hierarchical models and extensions. For instance, linear models have been developed to fit the logarithm of the variances (Foulley et al., 2004) in the univariate case. In the multivariate case, we think that BNs, not only GBNs, are appropriate candidates for further proposals. Recent publications appear in that perspective and we can cite Scutari et al. (2014) as an example. In this study, we showed that two-way structures can be introduced. Of course, there is no limitation to two ways, similar multi-factor approaches can be devised as well. The work by Murphy (1998) also uses BNs with constraints on the variances (so-called *tied variances*); but his focus is different from ours since he proposed equalities between conditional variances while we use the DAG structure itself to constraint the joint variance matrix on the set of nodes; the additional constraints defined of Table 1 are put on the conditional expectations.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

The need for modelling variance structure is all the more urgent now, that statisticians are facing more and more situations where p, the number of variables, is great while n, the number of statistical units to afford them, is small. BNs modelling seems a good way to take up the challenge.

We showed that, at least for GBN modelling, it was possible to introduce a known structure on the set of variables of interest, and that can lead to very effective results to obtain interpretable predictive formulae. The numerical experiments based on simulated data sets with a known structure showed that at least in some situations, some learning algorithms were sufficiently efficient to give our proposal a practical impact on data analysis. One of the advantages of the BN formulation is to allow non-statisticians, typically experts in some field, to contribute to model specification through the easy to understand DAG presentations. In our mind, such BNs must and can serve as thinking material for non-experts in BNs. In that respect, the availability of user-friendly and performing software is a prerequisite and we are happy to see that more and more R packages playing this role, are proposed: the most complete and versatile is BNLEARN (Scutari, 2010), but PCALG (Kalisch et al., 2012), DEAL (Bottcher and Dethlefsen., 2012) and IGRAPH (Csardi and Nepusz, 2006) are also worth mentioning.

Besides the introduction of structures on the set of the variables of interest, our study underlines the distinction between variables and covariates. One could think that the ideal model would be a model such that the targeted variables be conditionally independent to the covariates. That is all the covariation between them could be explained by external variables. With this respect some of the exhibited models, having a very small parametric dimension (for instance Model 12 of Table 6 with 3 instead of 36) are appealing.

Many more ideas could be proposed to achieve the goals we were interested in. Among them, the use of distributions other than Normal ones. Probably mathematical properties will be much more difficult to obtain, but the advantage would be to achieve a more realistic model specification. Advanced numerical tools already exist to undertake such an investigation. Among them, even if not originally devised for this purpose, are the BUGS software packages (Plummer, 2003 and Lunn et al., 2013). But also simpler approaches could also be worthwhile, like the use of transformations of the initial variables. More sophisticated constraints than the equalities could also be implemented. For instance, following again a two-way structure, some bilinear modelling could be thought about like those proposed to generalise additive models (Denis and Gower, 1996).

We do think that not only BNs are beautiful for mathematical aspects, they are also useful for plenty of applied questions within a classic statistics point of view. In particular, they allow to incorporate, in the statistical models used to the data sets to interpret, more prior knowledge about the phenomenon under study leading to more precise inferences and predictions. The proposals we made are going this way.

Acknowledgments

The authors acknowledge the relevant comments made by the reviewers and editors of the journal on a first version of the manuscript which lead to an improvement of the paper.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, 3rd edition.

Bang-Jensen, J. and Gutin, G. (2009). *Digraphs: Theory, Algorithms and Applications*. Springer, 2nd edition.

Bottcher, S. G. and Dethlefsen., C. (2012). *deal: Learning Bayesian Networks with Mixed Variables*. R package version 1.2-35.

Burnham, K. P. and Anderson, D. R. (2010). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Celeux, G., Marin, J.-M., and Robert, C. P. (2006). Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147(1):59–79.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Denis, J.-B. (2013). *rbmn: Handling Linear Gaussian Bayesian Networks*. R package version 0.9-2.

Denis, J.-B. and Gower, J. C. (1996). Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. *Applied Statistics*, 45(4):479–493.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer-Verlag.

Foulley, J.-L., Sorensen, D., Robert-Granié, C., and Bonaïti, B. (2004). Heteroskedasticity and structural models for variances. *Jour. Ind. Soc. Ag. Slatistics*, 57:64–70.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse Inverse Covariance Estimation With the Graphical Lasso. *Biostatistics*, 9:432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, N., Murphy, K., and Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In *Proceeding of the 14th Conference on Uncertainty and Artificial Intelligence (UAI'98)*, pages 139–147. Morgan Kaufmann.

Ghahramani, Z. (1997). *Learning Dynamic Bayesian Networks*. Number 1387 in Lecture Notes In Computer Science. Springer.

Graybill, F. A. (1976). *Theory and application of the linear model*. Duxbury Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition.

Jungnickel, D. (2013). *Graphs, Networks and Algorithms*. Springer, 4rth edition.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Korb, K. B. and Nicholson, A. E. (2011). *Bayesian Artificial Intelligence*. CRC press, 2nd edition.

Leray, P. (2006). *Réseaux bayésiens : apprentissage et modélisation de systèmes complexes*. PhD thesis, Université de Rouen. Habilitation à Diriger des Recherches.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The* BUGS *Book. A practical introduction to Bayesian analysis*. CRC press.

Miller, A. J. (2002). *Subset selection in regression*. Boca Raton: Chapman & Hall / CRC, 2d edition.

Murphy, K. P. (1998). Fitting a conditional gaussian distribution. Technical report.

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley. PhD dissertation.

Nagarajan, R., Scutari, M., and Lèbre, S. (2013). *Bayesian Networks in R with Applications in Systems Biology*. Springer.

Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O., and Becker, A. (2004). *Réseaux bayésiens*. Eyrolles, 2nd edition.

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Prentice Hall.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22.

Scutari, M. and Denis, J.-B. (2014). *Bayesian Networks with Examples in R*. Chapman & Hall. in print.

Scutari, M., Howell, P., Balding, D. J., and Mackay, I. (2014). Multiple quantitative trait analysis using bayesian networks. *Genetics*. in print.

Sedgewick, R. (2011). *Algorithms*. Addison-Wesley, 4th edition.

Tian, S., Mioche, L., Denis, J.-B., and Morio, B. (2013). A multivariate model for predicting segmental body composition. *British Journal of Nutrition*, 110(12):2260–70.

Timm, N. (2002). *Applied Multivariate Analysis*. Springer.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

© Société Française de Statistique et Société Mathématique de France (2014) ISSN: 2102-6238