Bayesian Network structural learning in multivariate time series

Maria Caterina Bramati and Flaminia Musella

Abstract Structural learning of Bayesian network is usually fulfilled by the expert knowledge, whenever available, or by some efficient algorithmic procedures. Despite the vast literature on structural learning, still little has been done specifically aimed at the multivariate time series modeling. We suggest a testing procedure able to learn the DAG structure whose vertex set only consists of the components of a stationary vector autoregressive VAR(p) model. The proposal procedure follows a constraint-based approach by using a test between blocks of variables. The class of tests proposed is based on multivariate ranks of distances and it is asymptotically distribution-free under very mild assumptions on the noise.

Key words: multivariate time series, multivariate ranks and signs, structural learning, VAR models

1 Introduction

Bayesian networks (BNs - Cowell et al. 1999) are graphical model able to model phenomena characterized by uncertainty. They are probabilistic models for representing the joint distribution of a set of variables by means of directed acyclic graphs (DAGs). A graph G is an object composed by a set of nodes, denoted by V and representing variables, and a set of (eventually directed) edges, denoted by E interlace pairs of nodes. A directed graph is said to be acyclic if it is not possible, following arrows direction, to define a path where the starting and the ending node coincide.

M.C. Bramati
Sapienza University of Rome, *Dpt. Methods and Models for Economics, Territory and Finance*, Via del Castro Laurenziano 9, I-00161 Rome, e-mail: mariacaterina.bramati@uniroma1.it

F. Musella
University of Roma Tre, *Dpt. of Economics*, Via S. D'Amico, 77 I-00145 Rome, e-mail: flaminia.musella@uniroma3.it

In a BN each node is associated with a (marginal or conditional) probability distribution function. When the dependence structure between variables is known, the DAG can be drawn on the basis of expert knowledge and the probability tables can be manually filled in. In many applications the dependence structure is unknown or partially known, and the network need to be automatically learnt from the data. The latter is the situation largely studied and discussed in the literature (Buntine 1994, Buntine 1996) and it is known as structural learning. The main approaches to structural learning are the *scoring and searching* and the *constraint-based*. The first one is based on searching, in a specific space, the model that has the best score, given a chosen metric (Cooper and Herskovits 1992; Heckerman 1995); the second focuses on the conditional independence relations revealed from the data and the network is drawn according to the test results (Spirtes et al. 2000). The paper discusses the opportunity to check conditional independence between blocks of variables to determine entire paths of the network by accounting for the temporal dimension. The paper is organized as follows. Section 2 briefly discusses the state of art of modeling temporal setting with graphical models and Bayesian networks. Section 3 concludes with the class of conditional test for checking independences.

2 Graphical modeling in temporal setting

The increasing interest in graphical models has been developed also in the time series modeling. Brillinger (1996) introduced the idea of using graphical models in time series analysis. The first attempt for modeling time series in graphical models framework has led to the chain graphs (Lynggaard and Walther, 1993). Some applications of the tool can be found in Reale and Tunnicliffe Wilson (2001), Dahlhaus and Eichler (2003). In these papers, the analysis is restricted on linear interdependencies among the variables. More recently, nonlinear parametric and nonparametric models have been discussed (e.g., Tong, 1993; Fan and Yao, 2003). A widespread approach is to consider the variables at different time slices and to represent them by separate nodes, with the consequence of graphs with many nodes. More recently, Eichler (1999) proposed an alternative approach for modeling multivariate stationary time series by using the concept of Granger causality expressed in terms of conditional independencies. The pictorial representation of the model is based on mixed graphs where: (1) each variable, i.e. each time series, is displayed by a single vertex; (2) Granger-causal relationships between pairs of variables is displayed by directed edges; (3) simultaneous dependence structure is represented by undirected edge.

The literature also focused on various approaches to incorporate the temporal dimension into BNs. The extension of BNs used to model time series is called dynamic Bayesian networks (DBNs - Webber and Jouffe, 2006; Dean and Kanazawa, 1989). In a DBN, state at time t is represented by a set of variables and depends on the states at previous time steps. A DBN is defined by one BN for every time slice. The time slices are repeated T-times. DBNs enable us to visualize intra-slice connec-

tions and inter-slice connections that incorporate conditional probabilities between variables from different time slices. This approach is a little benefit in the analysis of multivariate long time series.

In this research we focus on the relevant theoretical and computational problems related to the Bayesian Network structural learning in the multivariate time series setting. We adopt a constraint based approach to learn a BN where each node is a time series and arrows depict dependencies in a temporal dimension. The dependencies are checked with class of tests of conditional independence able to identify the temporal structure of the system. The problem is tackled in a Vector Autoregression framework where the temporal dimension is exploited in order to identify the direction of the edges in the BNs. This approach is quite new in the literature and would avoid undirected edges.

3 Block-wise conditional independence in time

Let $\mathbf{X}_t = \left((\mathbf{X}_t^{(1)})^T, (\mathbf{X}_t^{(2)})^T\right)^T$ be a $k \times 1$ vector of stochastic processes with $\mathbf{X}_t^{(1)}$ taking values in $\mathbb{R}^{k_1}$ and $\mathbf{X}_t^{(2)}$ in $\mathbb{R}^{k_2}$ ($k_1 + k_2 = k$), the Vector AutoRegressive model of order p, (denoted by VAR(p)) is

$$\mathbf{X}_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{X}_{t-i} + \boldsymbol{\varepsilon}_t \qquad t = 1, \ldots, N \tag{1}$$

where $\mathbf{A}_i$, $i = 1, \ldots, p$ are $k \times k$ matrices and $\boldsymbol{\varepsilon}_t = \left((\boldsymbol{\varepsilon}_t^{(1)})^T, (\boldsymbol{\varepsilon}_t^{(2)})^T\right)^T$ is a $k \times 1$ vector of noise.

Clearly, in a VAR model contemporaneous relationships among variables are not accounted for and are "hidden" in the contemporaneous correlation structure of the error terms. The structural VAR analysis is based on the attempt to give a sensible solution to this problem, based on the imposition of a set of restrictions. In this paper we address more deeply the problem of testing restrictions on the correlation structure of the noise.

Consider the following partition in sub-matrices of both the autocorrelation matrices, for $i = 1, \ldots, p$ and the invertible and symmetric $k \times k$ matrix $\mathbf{A}_0$

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{A}_i^{(11)} & \mathbf{A}_i^{(12)} \\ \mathbf{A}_i^{(21)} & \mathbf{A}_i^{(22)} \end{pmatrix} \qquad \mathbf{A}_0 = \begin{pmatrix} \mathbf{A}_0^{(11)} & \mathbf{A}_0^{(12)} \\ (\mathbf{A}_0^{(12)})^T & \mathbf{A}_0^{(22)}, \end{pmatrix}$$

which represents a perturbation of a block spherical noise $\boldsymbol{\eta}_t = \left((\boldsymbol{\eta}_t^{(1)})^T, (\boldsymbol{\eta}_t^{(2)})^T\right)^T$, i.e. $\boldsymbol{\varepsilon}_t := \mathbf{A}_0 \boldsymbol{\eta}_t$, where $\mathbf{A}_0^{(11)}$ and $\mathbf{A}_0^{(22)}$ are nonsingular symmetric $k_1 \times k_1$ and $k_2 \times k_2$ matrices, respectively, and $\mathbf{A}_0^{(12)}$ is a $k_1 \times k_2$ matrix. Without loss of generality we assume that $[\mathbf{A}_0]_{11} = 1$.

From the definition of instantaneous conditional independence, which in the Gaussian case is equivalent to conditional non correlation, testing for contemporaneous non correlation between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is testing the null hypothesis

$$\mathscr{H}_0: \quad \mathbf{A}_0^{(12)} = \mathbf{0}. \tag{2}$$

Two necessary requirements are stationarity of the stochastic process $\mathbf{X}_t$ in model (1) and finite second-order moments. Moreover, we consider $\boldsymbol{\eta}_t = \big((\boldsymbol{\eta}_t^{(1)})^T, (\boldsymbol{\eta}_t^{(2)})^T\big)^T$ a marginally spherical noise, i.e. with density

$$\underline{f}(\mathbf{z}; f_1, f_2) := c_{k,f_1,f_2} f_1(\|\mathbf{z}^{(1)}\|) f_2(\|\mathbf{z}^{(2)}\|)$$

where $c_{k,f_1,f_2} \in \mathbb{R}$, $\mathbf{z}^{(1)} \in \mathbb{R}^{k_1}$, $\mathbf{z}^{(2)} \in \mathbb{R}^{k_2}$ and $f_1 : \mathbb{R}_0^+ \to \mathbb{R}^+$, $f_2 : \mathbb{R}_0^+ \to \mathbb{R}^+$ are a.e. strictly positive are the marginal *radial densities* of $\boldsymbol{\eta}_t^{(1)}$ and $\boldsymbol{\eta}_t^{(2)}$, respectively, and are assumed to have finite *radial Fisher information for location and scale*.

Under the null hypothesis of non correlation, $\boldsymbol{\varepsilon}_t := \mathbf{A}_0 \boldsymbol{\eta}_t$ is a marginally elliptical error term.

In synthesis, we require in (A1) the stationarity of the VAR(p) model considered in (1), and we assume (A2) that under the null the innovations are marginally elliptical.

Let

$$\boldsymbol{\theta}_I := (\mathrm{vec}^T \mathbf{A}_1, \ldots, \mathrm{vec}^T \mathbf{A}_p)^T, \quad \boldsymbol{\theta}_{II} := ((\mathrm{vech}_0^* \mathbf{A_0})^T)^T$$

and $\boldsymbol{\vartheta} := (\boldsymbol{\theta}_I^T, \boldsymbol{\theta}_{II}^T)^T$ be respectively the pk^2, the $k(k+1)/2-1$ and the $pk^2 + k(k+1)/2 - 1$-dimensional vectors of unknown parameters corresponding to the model in (1), where the vech_0^* operator is such that, for any $k \times k$ matrix $\mathbf{A}$,

$$\mathrm{vech}_0^* \mathbf{A} := (\mathrm{vech}_0^T \mathbf{A}^{(11)}, \mathrm{vec}^T \mathbf{A}^{(21)}, \mathrm{vech}^T \mathbf{A}^{(22)})^T$$

with the usual definitions of vec, vech and vech_0 (the last operator corresponds to the vech operator, having dropped the top-left entry of the matrix $\mathbf{A}$).

Vector $\boldsymbol{\theta}_I$ represents a vector of nuisance parameters in the testing problem. When $\boldsymbol{\theta}_I$ is left unspecified, the multivariate signs and ranks-based test statistics proposed in this paper is computed using any arbitrary estimators $\hat{\boldsymbol{\theta}}_I = \hat{\boldsymbol{\theta}}_I^{(n)}$ satisfying assumptions previously specified. Then the test still allows to achieve local asymptotic optimality in the Le Cam sense, under adequate choices of the score functions.

Let $\mathbf{U}_t := ((\mathbf{U}_t^{(1)})^T, (\mathbf{U}_t^{(2)})^T))^T$, where $\mathbf{U}_t^{(1)} := \mathbf{z}_t^{(1)}/d_t^{(1)}$ and $\mathbf{U}_t^{(2)} := \mathbf{z}_t^{(2)}/d_t^{(2)}$ are the unit-norm vectors which under $P^{(N)}_{\boldsymbol{\vartheta};f_1,f_2}$ are i.i.d. and uniformly distributed over the k_1 and k_2-dimensional unit sphere. Let $R_t^{(j)}$ be the rank of $d_t^{(j)}$ among $(d_1^{(j)}, \ldots, d_N^{(j)})$, $j = 1, 2$, then define the sign and rank-based test statistics

$$\underset{\sim}{\mathscr{W}}^{(N)}_{f_1,f_2}(\boldsymbol{\vartheta}) := \underset{\sim}{\mathbf{T}}^{(N)T}_{f_1,f_2}(\boldsymbol{\vartheta}) \underset{\sim}{\mathbf{C}}^{-1}_{f_1,f_2} \underset{\sim}{\mathbf{T}}^{(N)}_{f_1,f_2}(\boldsymbol{\vartheta}) \tag{3}$$

with

$$\underset{\sim f_1,f_2}{\boldsymbol{T}}^{(N)}(\boldsymbol{\vartheta}) := \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\underset{\sim t,f_1,f_2}{\mathbf{B}}\,\mathrm{vec}\,\mathbf{U}_t^{(1)}\mathbf{U}_t^{(2)T} \tag{4}$$

$$\begin{aligned}\underset{\sim t,f_1,f_2}{\mathbf{B}} &:= J_1^{(1)}\left(\frac{R_t^{(1)}}{N+1}\right)J_0^{(2)}\left(\frac{R_t^{(2)}}{N+1}\right)(\mathbf{I}_{k_2}\otimes\mathbf{A}_0^{(11)})^{-1}\\ &\quad + J_1^{(2)}\left(\frac{R_t^{(2)}}{N+1}\right)J_0^{(1)}\left(\frac{R_t^{(1)}}{N+1}\right)(\mathbf{A}_0^{(22)}\otimes\mathbf{I}_{k_1})^{-1},\end{aligned}$$

and

$$\underset{\sim f_1,f_2}{\mathbf{C}} := \frac{1}{k_1k_2}E_{f_1,f_2}[\underset{\sim t,f_1,f_2}{\mathbf{B}}\,\underset{\sim t,f_1,f_2}{\mathbf{B}^T}],$$

where the score functions defined as $J_1^{(1)}(u) := \varphi_{f_1}\circ\tilde{F}_1^{-1}(u)$, $J_1^{(2)}(u) := \varphi_{f_2}\circ\tilde{F}_2^{-1}(u)$, $J_0^{(1)}(u) := \tilde{F}_1^{-1}(u)$ and $J_0^{(2)}(u) := \tilde{F}_2^{-1}(u)$, satisfy assumptions on the score function previously mentioned. Functions $\tilde{F}_1(\cdot)$ and $\tilde{F}_2(\cdot)$ are the distribution functions of the moduli of the sphericized residuals under marginal radial densities $f_1(\cdot)$ and $f_2(\cdot)$.

Let $J_1(u) = K_1(u)$, $J_0(u) = K_0(u)$ be any continuous differentiable monotone increasing square-integrable score functions then the general score version of the test statistics (3) is

$$\underset{\sim J}{\mathscr{W}}^{(N)}(\boldsymbol{\vartheta}) := \underset{\sim J}{\mathbf{T}}^{(N)T}(\boldsymbol{\vartheta})\underset{\sim J}{\mathbf{C}}^{-1}\underset{\sim J}{\mathbf{T}}^{(N)}(\boldsymbol{\vartheta}) \tag{5}$$

with

$$\underset{\sim J}{\boldsymbol{T}}^{(N)}(\boldsymbol{\vartheta}) := \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\underset{\sim t}{\mathbf{B}}\,\mathrm{vec}\,\mathbf{U}_t^{(1)}\mathbf{U}_t^{(2)T} \tag{6}$$

$$\underset{\sim J}{\mathbf{C}} := \frac{1}{k_1k_2}E_{f_1,f_2}[\underset{\sim t}{\mathbf{B}}\,\underset{\sim t}{\mathbf{B}^T}].$$

In particular, when the score functions $K_1^{(j)}(\cdot)$ and $K_0^{(j)}(\cdot)$ are equal to a function $K^j(\cdot)$, $j = 1,2$, the general score version of the test statistics boils down to

$$\underset{\sim J}{\mathscr{W}} = \frac{k_1k_2}{E[K^{(1)2}]E[K^{(2)2}]}\left\|\frac{1}{\sqrt{N}}\sum_{t=1}^{N}K^{(1)}\left(\frac{R_t^{(1)}}{N+1}\right)K^{(2)}\left(\frac{R_t^{(2)}}{N+1}\right)\mathrm{vec}\,\mathbf{U}_t^{(1)}\mathbf{U}_t^{(2)T}\right\|^2.$$

Then, under some mild assumptions, such as stationarity, finite second order moments, *constraintness*, *root-n consistency*, *local asymptotic discreteness*, *affine equivariance* and continuous differentiable monotone increasing square-integrable score functions[1], the test statistics $\underset{\sim f_1,f_2}{\mathscr{W}}^{(N)}(\boldsymbol{\vartheta})$ is block affine-invariant and asymptotically invariant with respect to the group of continuous monotone marginal radial

[1] see Bramati (2013)

transformations; is asymptotically chi-square with k_1k_2 degrees of freedom under $\bigcup_{\boldsymbol{\theta}_{II}\in\mathscr{M}(\mathbf{Q})}\bigcup_{\boldsymbol{\theta}_I}\bigcup_{g_1,g_2}P^{(N)}_{\boldsymbol{\theta}_{II};\boldsymbol{\theta}_I,g_1,g_2}$ (so that $\phi^{(N)}_{\sim J}$ has asymptotic level α); is asymptotically noncentral chi-square with k_1k_2 degrees of freedom and noncentrality parameter

$$\psi^2(\boldsymbol{\tau}_{II},\boldsymbol{\vartheta}) = \|\mathbf{C}_{J,g_1,g_2}\,\mathrm{vec}\,\mathbf{H}_{(12)}\|^2_{\mathbf{C}_J},$$

under $P^{(N)}_{\boldsymbol{\theta}_{II}+N^{-1/2}\boldsymbol{\tau}_{II};\boldsymbol{\theta}_I,g_1,g_2}$, $\boldsymbol{\tau}_{II}\notin\mathscr{M}(\mathbf{Q})$.

This research aims at combining uncertainty and temporal dimension in Bayesian Networks. The temporal dimension is generally exploited in order to identify the direction of the edges in the BNs and thus avoiding an a priori given ordering. This is done for understanding the relations of the system, with the simple idea that an effect cannot precede its cause in time. The idea is to learn the structure of the BNs using constraint based approach based on block-wise conditional independence tests in time.

4 References

Bramati, M.C. (2013). A class of optimal tests for contemporaneous noncausality in VAR models. *Journal of Time Series Analysis*, vol. 34, pp. 330-344.

Cooper, G. and Herskovits, E. (1992). A Bayesian method for constructing Bayesian belief networks from databases. *Machine Learning*, **9**, pp. 309- 347.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Dahlhaus, R. and Eichler, M. (2003). Causality and graphical models in time series analysis. In P. Green, N. Hjort and S. Richardson (eds), *Highly structured stochastic systems*, University Press, Oxford.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation, *Artificial Intelligence* **93**, pp. 1-27.

Eichler, M. (1999). *Graphical Models in Time Series Analysis*. Doctoral thesis, University of Heidelberg, Germany.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

Heckerman, D. (1995). *A tutorial on Learning with Bayesian Networks*. Technical Report MSR-TR-95-06, Microsoft Research.

Lynggaard, H. and Walther, K. H. (1993). *Dynamic modelling with mixed graphical association models*. Master's thesis, Aalborg

Reale, M. and Tunnicliffe Wilson, G. (2001). Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications*, **10**, pp. 49-65.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*, (2nd edn). MIT Press, Cambridge, Massachusetts.

Tong, H. (1993). *Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.