CAUSALITY IN VECTOR AUTOREGRESSIONS A BAYESIAN GRAPHICAL MODELLING APPROACH

JUKKA CORANDER AND MATTIAS VILLANI

ABSTRACT. The notion of causality in multiple time series is addressed from a Bayesian graphical modelling perspective in the class of vector autoregressive (VAR) processes. Due to the very large number of graph structures that may be considered, simulation based inference, such as Markov chain Monte Carlo, is not feasible. Therefore, we derive an approximate joint posterior distribution of the number of lags in the autoregression and the causality structure represented by graphs using a fractional Bayes approach. Some properties of the approximation are derived and the analysis is illustrated on a four-dimensional macroeconomic system and five-dimensional air pollution data.

1. INTRODUCTION

The vector autoregression (VAR) can be written as

$$x_t = \sum_{i=1}^{k} \Pi_i x_{t-i} + \varepsilon_t, \quad t = 1, ..., n, \tag{1.1}$$

where x_t is a p-dimensional vector of time series observations at time t, Π_i are $p \times p$ coefficient matrices determining the dynamics of the system and $\varepsilon_1, ..., \varepsilon_n \overset{iid}{\sim} N_p(0, \Sigma)$. Deterministic variables can be added to the model, leading only to trivial modifications of the results obtained here.

The statistical properties of the VAR model are by now well explored, see *e.g.* Lütkepohl (1993). The number of parameters in the VAR model is typically very large and it is complicated to investigate the dynamic relations between the time series simply by looking at estimates of $\Pi_1, .., \Pi_k$ and Σ. Several tools have been suggested to aid in the interpretation of VAR models, most notably the impulse response functions methodology introduced by Sims (1980), and causality tests (Granger, 1969; Sims, 1972).

Our focus here is on causality relations modelled by mathematical *graphs*. *Graphical models*, *i.e.* models which use mathematical graphs to represent multivariate relations, have become widely known and used statistical tools in cross-sectional data analysis, see *e.g.* Whittaker (1990), Lauritzen (1996) and Wermuth (1998). A typical graph consists of a set of vertices representing the variables and a set of edges between these vertices. The presence of an edge between a pair of vertices means that the variables are, in some sense, related. In multivariate cross-sectional data the graph usually represents the conditional independence structure of the system and, as independence is a symmetric relation, the edges are undirected. The time dimension of a process makes it possible

Key words and phrases. Causality, Fractional Bayes, graphical models, lag length selection, vector autoregression.

The authors are grateful to Prof. H. Karrasch, Geographisches Institut, for the air pollution data set. The second author gratefully acknowledges financial support from the Swedish Council of Research in Humanities and Social Sciences (HSFR), grant no. F0582/1999 and the Swedish Research Council (Vetenskapsrådet) grant no. 412-2002-1007.

to consider directed flow, or causality, and it is therefore natural to set up a graph for a time series system with this notion defining the relations between variables, i.e. the presence of an edge from one variable to another implies a causal relation in the direction indicated by the edge.

A majority of the research concerning graphical models has concentrated on modelling cross-sectional multivariate observations and only recently have more systematic efforts been made to utilize graph concepts in the context of stochastic processes (Eichler, 1999, Dahlhaus, 2000, Dahlhaus and Eichler, 2001). These works have focused on the development of fundamental probabilistic properties of graphical models for multivariate time series in a general context, while leaving the inference aspects more open for further research. In particular, there have not been any attempts to use Bayesian methods. As inference in graphical models is essentially a model determination problem, a field where Bayesian methods have dominated during the last decade, the lack of Bayesian research in this area is surprising. In the cross-sectional data domain, on the other hand, the potential of Bayesian analysis has been rapidly recognized, and there is a large literature on Bayesian inference for graphical models of cross-sectional data, see *e.g.* Dawid and Lauritzen (1993), Madigan and Raftery (1994), Madigan and York (1995), Dellaportas and Forster (1999), Giudici and Green (1999), and Corander (2001). Convincing arguments for a Bayesian approach in a more general setting may be found in the many books on the subject, see *e.g.* Bernardo and Smith (1994) and the references therein.

We concentrate here on the class of VAR-processes for mainly two reasons: first, the VAR process is widely used in applied time series analysis, and second, statistical inference proves to be tractable for this class of models. In contrast to the traditional graphical modelling of cross-sectional data, the widely used Markov chain Monte Carlo (MCMC) techniques do not provide a practical solution for the graphical models considered here, mainly due to the much larger space of possible graph structures. To make inference jointly about the causality structure and the lag length of the process, we use the fractional Bayes approach of O'Hagan (1995), which has proven to be well suited for multiple time series analysis (Villani, 2001a).

The paper is organized as follows. In Section 2 we discuss various types of graphical models together with necessary concepts and results from graph theory. In Section 3 the joint posterior distribution for the lag length and the causality structure is derived. The inference procedure is illustrated numerically in Section 4, and some concluding remarks are given in the final section.

2. Graphical VAR models

2.1. **Independence graphs for cross-sectional data.** We begin by presenting the traditional independence graph for cross-sectional data together with a review of the necessary concepts from graph theory, see Whittaker (1990) and Lauritzen (1996) for a detailed treatment. A visual representation of conditional independencies is given by an *independence graph*, G, which consists of two sets, a set of *vertices*, $V = \{1, ..., p\}$, and a set of *edges* $E \subseteq \{(i, j) \in V \times V\}$, connecting some of the vertices. The variables are represented by vertices in the graph and the absence of an edge between a pair of vertices implies that the corresponding variables are independent conditional on the other variables in the system. A precise definition is as follows.

Definition 1. A conditional independence graph $G = (V, E)$ on a set of random variables $X_1, ..., X_p$ is defined by the relation

$$(i, j) \notin E \Longleftrightarrow X_i \perp X_j \mid X_{V\setminus\{i,j\}},$$

where $\perp$ denotes independence and $V\setminus a = \{v \in V : v \notin a\}$ for any $a \subseteq V$.

A graph is called *undirected* if $(i, j) \in E \Leftrightarrow (j, i) \in E$, i.e. if the presence of an edge between vertex i and j implies that there is also an edge in the reverse direction. As independence is a symmetric relation, independence graphs are undirected. A subset $a \subseteq V$ of vertices *separates* two vertices i and j if every path from i to j crosses at least one vertex in a. A subset $a \subseteq V$ separates two subsets $b, c \subseteq V$ if it separates every pair of vertices $i \in b$ and $j \in c$. The induced *subgraph* of $a \subseteq V$, G_a, is obtained by removing from V all vertices not in a together with all edges which do not join two vertices in a. A graph is *complete* if it has maximum number of edges. A subset $a \subseteq V$ is called a *clique* if it is *maximally complete*, i.e. if its induced subgraph is complete but the induced subgraph of any extension of a is incomplete.

An undirected graph is said to be *decomposable* or *triangulated* if it has no chordless cycles of length larger than three, which implies that G can be broken into a sequence of its cliques by certain basic operations. From this sequence, the *separators* of a decomposable graph can be obtained as intersections of successive cliques. The joint density of a set of random variables $X_1, ..., X_p$ with decomposable independence graph factorizes as follows (Whittaker, 1990; Lauritzen, 1996)

$$p(x_1, ..., x_p) = \frac{\prod_{c\in\mathcal{C}(G)} p(x^{(c)})}{\prod_{s\in\mathcal{S}(G)} p(x^{(s)})}, \tag{2.1}$$

where $\mathcal{C}(G)$ and $\mathcal{S}(G)$ are the sets of cliques and separators, respectively, in the graph, and $x^{(c)}$ and $x^{(s)}$ are subsets of variables in clique c and separator s, respectively. The decomposition in (2.1), realizable by the assumed decomposability of the graph, splits the density, in a sense, into conditionally independent pieces. It is precisely this property which makes the class of decomposable graphical models especially attractive from a Bayesian point of view, see e.g. Dawid and Lauritzen (1993), Madigan and Raftery (1994), Madigan and York (1995), Giudici and Green (1999).

For independent Gaussian observations with covariance matrix Σ it is well known that the independence graph can be characterized by the relation $(i, j) \notin E \Leftrightarrow \Omega(i, j) = 0$, where $\Omega(i, j)$ is the (i, j)th element of $\Omega = \Sigma^{-1}$. By assuming that E is decomposable, traditional methods (see the references in the previous paragraph) may be used directly on the ε_t-sequence in (1.1), with a missing edge between node i and j signifying that the two time series X_i and X_j are *contemporaneously independent* conditional on the other variables. Contemporaneous independence is of course only part of the dependence structure for time series and the next section describes a graph structure developed by Dahlhaus (2000) which takes the time dimension of time series data into account.

2.2. **Partial correlation graphs for time-series data.** The undirected *partial correlation graph* for a multivariate process may be defined as follows.

Definition 2. A *partial correlation graph* $G = (V, E)$ on a set of p stationary time series $X_1, ..., X_p$ is defined by the relation

$$(i, j) \notin E \Longleftrightarrow r_{ij}(\lambda) = 0, \ \forall \ \lambda \in (-\pi, \pi),$$

where $r_{ij}(\lambda)$ denotes partial spectral coherence of X_i and X_j (Brillinger, 1981).

The partial correlation graph was first defined by Dahlhaus (2000), who actually proved our definition using an alternative definition based on orthogonality of the residuals of X_i and X_j with the effect of $X_{V\setminus\{i,j\}}$ and linear trends removed.

For Gaussian processes, and for the VAR process (1.1) in particular, the partial correlation graph has a conditional independence interpretation and may rightly be called a generalized independence graph. However, the partial correlation graph need not be as informative as one would expect; this is illustrated in the next section, after the introduction of yet another graph type: the causality graph.

2.3. **Causality graphs for time-series data.** A relatively detailed representation of the dependence structure is obtained by supplementing the decomposable independence graph on ε_t-sequence in (1.1) with some notion of directed flow or causality. Here we follow Eichler (1999) and Dahlhaus and Eichler (2001) in the use of Granger causality (Granger, 1969) and in the restriction of attention to stationary time series. Loosely defined, the time series X_j *Granger-causes* another series X_i if and only if the addition of X_j to the (existing) predictor set improves the predictions of X_i in a mean square sense; see Granger (1969) for an exact definition. It is clear that any causality relation is directed; the presence of a directed edge between X_i and X_j does not imply the existence of an edge in the opposite direction.

Graphs consisting of both directed and undirected edges are usually referred to as *mixed*. The edge set E of a mixed graph G is the union $E = E_1 \cup E_2$, where edges in E_1 and E_2 are directed and undirected, respectively. Such mixed graphs will be denoted by $G^{\rightleftharpoons} = (V, E_1, E_2)$, whereas undirected graphs will be denoted by $G^{\sim}$ to make a clear distinction between the two graph types. There are $2^{\binom{p}{2}}$ undirected and $2^{3\binom{p}{2}}$ mixed graphs, respectively, for a set of p vertices.

Definition 2.4 in Dahlhaus and Eichler (2001), reinterpreted in our setting, reads:

Definition 3. A causality graph $G^{\rightleftharpoons} = (V, E_1, E_2)$ on a set of stationary time series $X_1, ..., X_p$ is defined by the relations

$(i,j) \notin E_1 \Longleftrightarrow X_i$ is non-causal for X_j conditional on $X_{V\setminus\{i,j\}}$

$(i,j) \notin E_2 \Longleftrightarrow X_i$ and X_j are contemporaneously independent conditional on $X_{V\setminus\{i,j\}}$

If, in addition, E_2 defines a decomposable graph, then $X_1, ..., X_p$ is said to have a decomposable causality graph.

Lemma 1. If $X_1, ..., X_p$ follows a p-dimensional VAR process with causality graph $G^{\rightleftharpoons} = (V, E_1, E_2)$, then the following holds

$$(i,j) \notin E_1 \Longleftrightarrow \Pi_l(i,j) = 0, \text{ for all } l = 1, ..., k,$$
$$(i,j) \notin E_2 \Longleftrightarrow \Omega(i,j) = 0,$$

where $\Pi_l(i,j)$ is the (i,j)th element of Π_l and $\Omega = \Sigma^{-1}$.

Proof. Follows directly from Corollary 2.2.1 in Lütkepohl (1993) and Corollary 6.3.4 in Whittaker (1990). ∎

A VAR process satisfying the conditions in Lemma 1 will be called a graphical causal VAR model, or GCVAR($G^{\rightleftharpoons}, k$) for short.

As shown in Dahlhaus and Eichler (2001), the graph theoretic concept of *moralization* which refers to conversion of the directed edges into undirected ones and addition of eventual further undirected edges according to certain rules, may be used for deriving

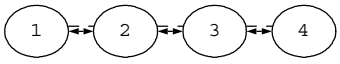

FIGURE 1. Causality graph of the VAR process specified in Eqn. 2.2.

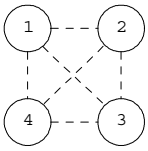

FIGURE 2. Partial correlation graph of the VAR process specified in Eqn. 2.2.

the edges of the partial correlation graph from the causality graph (see also Andersson et al., 2001). Using this relationship between the two types of graphs, we now illustrate that partial correlation graphs may disregard important aspects of dependence structure of a multivariate process.

Let $G^{\rightleftharpoons} = (V, E_1, E_2)$ be specified according to the graph in Figure 1. In Figure 1, and in the other illustrations we have used double-headed arrows to indicate the presence of a directed edge in both directions within a pair of vertices. Also, to clearly distinguish between directed and undirected edges, the latter are shown as dashed lines. Assuming further the lag length $k = 1$, the GCVAR$(G^{\rightleftharpoons}, k)$ model is given by

$$\Pi_1 = \begin{pmatrix} \pi_{11} & \pi_{12} & 0 & 0 \\ \pi_{21} & \pi_{22} & \pi_{23} & 0 \\ 0 & \pi_{32} & \pi_{33} & \pi_{34} \\ 0 & 0 & \pi_{43} & \pi_{44} \end{pmatrix} \quad \text{and} \quad \Omega = \begin{pmatrix} \omega_{11} & & & \\ \omega_{21} & \omega_{22} & & \\ 0 & \omega_{32} & \omega_{33} & \\ 0 & 0 & \omega_{43} & \omega_{44} \end{pmatrix}. \tag{2.2}$$

The conditional independence graph for this model equals the complete graph (Figure 2) and therefore carries no information regarding the rather pronounced variable order revealed by the causality graph.

3. Bayesian model assessment in GCVARs

The unknown quantities of a GCVAR$(G^{\rightleftharpoons}, k)$ process are: the underlying causality graph, $G^{\rightleftharpoons}$, the number of lags, k, and, conditional on a $(G^{\rightleftharpoons}, k)$-pair, the elements in $\Pi_1, ..., \Pi_k$ and Σ which are unrestricted under $G^{\rightleftharpoons}$. For notational simplicity we use θ as a shorthand for the free parameters of the GCVAR process; a more correct, but cumbersome, notation would be $\theta(G^{\rightleftharpoons}, k)$. Clearly, inference on $G^{\rightleftharpoons}$ and k must be settled before θ can be considered.

The main purpose of this paper is to derive the joint posterior distribution of $(G^{\rightleftharpoons}, k)$ conditional on the observed time series, which are for simplicity denoted by X. Let K be an *a priori* specified upper bound for the value of k. Using Bayes rule, the joint posterior distribution of $G^{\rightleftharpoons}$ and k then reads

$$\pi(G^{\rightleftharpoons}, k|X) = \frac{m(G^{\rightleftharpoons}, k, X)\pi(G^{\rightleftharpoons}, k)}{\sum_{G \in \mathcal{G}} \sum_{k=0}^{K} m(G^{\rightleftharpoons}, k, X)\pi(G^{\rightleftharpoons}, k)}, \tag{3.1}$$

where $\pi(G^{\overrightarrow{\sim}}, k)$ is the joint prior of $G^{\overrightarrow{\sim}}$ and k, $\mathcal{G}$ is the class of models under consideration, and

$$m(G^{\overrightarrow{\sim}}, k, X) = \int L(X|\theta, G^{\overrightarrow{\sim}}, k)\pi(\theta|G^{\overrightarrow{\sim}}, k)d\theta,$$

is the marginal likelihood of the observed time series X, where $L(X|\theta, G^{\overrightarrow{\sim}}, k)$ is the usual likelihood function under model $(G^{\overrightarrow{\sim}}, k)$ with parameters θ and $\pi(\theta|G^{\overrightarrow{\sim}}, k)$ is the prior distribution of θ. The joint prior of k and $G^{\overrightarrow{\sim}}$ is over a discrete set and can be chosen in many ways, e.g., if there is no reason for favoring any particular graph in $\mathcal{G}$ *a priori* one may use

$$\pi(G^{\overrightarrow{\sim}}, k) = \pi(k)/|\mathcal{G}|, \tag{3.2}$$

where $\pi(k)$ is some discrete distribution over the integers $k = 0, 1, ..., K$. The symbol $|\cdot|$ will be used to represent both the cardinality of a set and the determinant of a matrix.

In the sequel, let $\mathcal{G}$ denote the class of decomposable GCVAR$(G^{\overrightarrow{\sim}}, k)$ models. As θ varies with $G^{\overrightarrow{\sim}}$ and k, elicitation of subjective priors for θ can be a difficult and time-consuming task. In many problems one can find a relatively rich parametric family of distributions which both leads to a tractable posterior distribution and at the same is quite easily specified via a few hyperparameters. This is not the case for the Gaussian graphical models, including the model here, see, *e.g.*, the discussion in Giudici and Green (1999). Here we use a model-based reference prior, which can be utilized without further subjective assessment of prior hyper parameters. Using Σ_a to denote the $|a| \times |a|$ submatrix of Σ formed by the rows and columns of Σ corresponding to the set a, the prior is of the form

$$\pi(\theta|G^{\overrightarrow{\sim}}, k) \propto \frac{\prod_{c\in\mathcal{C}(G^{\sim})} |\Sigma_c|^{-(|c|+1)/2}}{\prod_{s\in\mathcal{S}(G^{\sim})} |\Sigma_s|^{-(|s|+1)/2}}. \tag{3.3}$$

The prior in (3.3) is obtained as a limit of a hyper inverse Wishart distribution (see Dawid and Lauritzen, 1993, or Giudici and Green, 1999) for Σ with degrees of freedom approaching zero (Geisser, 1965).

The prior in (3.3) is improper and is therefore not directly usable for deriving the joint posterior of $G^{\overrightarrow{\sim}}$ and k, as explained by *e.g.* O'Hagan (1995). A solution to this problem is to sacrifice a small part of the sample in updating the improper prior to a proper posterior and subsequently use this posterior as a new prior for the remaining observations. To avoid the arbitrary choice of training observations, O'Hagan (1995, 1997) suggested that the likelihood of the training sample could be approximated by a fraction of the likelihood for the whole sample, thereby replacing the choice of specific training observations to the much easier choice of training fraction. Thus, in the fractional Bayes approach to model inference the marginal likelihood in (3.1) is replaced by the *fractional marginal likelihood* (FML)

$$m_b(G^{\overrightarrow{\sim}}, k, X) = \frac{\int L(X|\theta, G^{\overrightarrow{\sim}}, k)\pi(\theta|G^{\overrightarrow{\sim}}, k)d\theta}{\int L(X|\theta, G^{\overrightarrow{\sim}}, k)^b\pi(\theta|G^{\overrightarrow{\sim}}, k)d\theta}, \tag{3.4}$$

where $0 < b < 1$ is the fraction of the data used to convert the improper prior to a proper posterior.

The fractional Bayes approach was in Villani (2001a) shown to produce favorable results in inference about k in the traditional VAR setting, compared to earlier established default solutions. As in Villani (2001a), it will be assumed here that b is *minimal*, *i.e.*

$b = m/n$, where m is the smallest number of observations yielding a proper posterior under the *largest* model in $\mathcal{G}$ (which is the complete graph).

The following lemma, which specifies the FML for a subclass of $\mathcal{G}$, will be used as a starting point for the FML in all of $\mathcal{G}$. In the lemma, X denotes the $n \times p$ matrix of row-stacked observations on the p time series, X_c is the $n \times |c|$ submatrix of X consisting of the columns corresponding to the variables in clique c and $X_{c(-l)}$ is X_c lagged l time periods.

Lemma 2. Assuming the prior (3.3), the fractional marginal likelihood of a p-dimensional GCVAR($G^{\rightleftarrows}, k$) process with $\mathcal{S}(G^{\sim}) = \varnothing$ and $G_c^{\rightleftarrows}$ complete for each $c \in C(G^{\sim})$ is

$$m_b(G^{\rightleftarrows}, k, X) = \prod_{c \in \mathcal{C}(G^{\sim})} \frac{\Gamma_{|c|}(n)}{\Gamma_{|c|}(m)} \left|\hat{\Sigma}_c\right|^{-(n-m)/2},$$

where $\Gamma_{|c|}(l) = \prod_{j=1}^{|c|} \Gamma[(l - k\,|c| - j + 1)/2]$, $\Gamma(\cdot)$ is the ordinary gamma function,

$$\begin{aligned}
\hat{\Sigma}_c &= n^{-1}(X_c - Z_c\hat{\Lambda}_c)'(X_c - Z_c\hat{\Lambda}_c) \\
Z_c &= (X_{c(-1)}, ..., X_{c(-k)}), \\
\hat{\Lambda}_c &= (Z_c'Z_c)^{-1}Z_c'X_c,
\end{aligned}$$

a multiplicative constant, common to all graphs and k, has been discarded, and $m = p(K + 1)$ yields the minimal training fraction $b = m/n$.

Proof. The lemma is a straightforward extension of Theorem 3.1 in Villani (2001a), applied to independent clique processes. ∎

Unfortunately, the marginal likelihood does not factorize under a general GCVAR($G^{\rightleftarrows}, k$) model with overlapping cliques in $G^{\sim}$ as it does for ordinary Gaussian graphical models (see Dawid and Lauritzen, 1993, or Giudici and Green, 1999). This is due to the discordance between the inference for a separator $s \in \mathcal{S}(G^{\sim})$ *within* a clique $c \in \mathcal{C}(G^{\sim})$ and the inference where all cliques in $\mathcal{C}(G^{\sim})$ containing s are considered jointly. For a detailed discussion of this model marginalization issue at a general level, see Lauritzen (1996). To enable at least an approximate solution to the model assessment problem within the class $\mathcal{G}$ with a non-empty separator set, we follow Corander and Villani (2001) and propose an approximation of the FML.

In the sequel, let the *subset specific in-degree* of vertex i in $a \subseteq V$ be defined as the difference $d_a(i)$ between cardinalities $|j \in V : (j,i) \in E_1|$-$|\{j \in V \backslash a : (i,j) \in E_2\} \cap \{(j,i) \in E_1\}|$. That is, $d_a(i)$ counts the total number of directed edges to i minus the number of directed edges to i from vertices j outside a such that j is adjacent to i in $G^{\sim} = (V, E_2)$. Notice that the sum of the subset specific in-degrees $\sum_{j=1}^{|a|} d_a(j)$ equals the total number directed edges in $G_a^{\rightleftarrows}$. Furthermore, $\hat{\Sigma}(G^{\rightleftarrows}, k)$ denotes the maximum likelihood (ML) estimate of Σ under the restrictions given by $G^{\rightleftarrows}$ and k. To simplify notation, we will not write out explicitly which model the parameters and their estimates refer to when this is evident from the discussion.

Definition 4. Assuming the prior (3.3), the approximate fractional marginal likelihood of a p-dimensional GCVAR$(G^{\rightleftharpoons}, k)$ is

$$m_b(G^{\rightleftharpoons}, k, X) = \frac{\prod_{c\in\mathcal{C}(G^{\sim})} \frac{\Gamma^*_{|c|}(n)}{\Gamma^*_{|c|}(m)} \left|\hat{\Sigma}_c\right|^{-(n-m)/2}}{\prod_{s\in\mathcal{S}(G^{\sim})} \frac{\Gamma^*_{|s|}(n)}{\Gamma^*_{|s|}(m)} \left|\hat{\Sigma}_s\right|^{-(n-m)/2}},$$

where $\hat{\Sigma}_a$ is the submatrix of the maximum likelihood estimate of Σ under the restrictions given by $G^{\rightleftharpoons}$ formed by the rows and columns corresponding to a and $\Gamma^*_{|a|}(l) = \prod_{j=1}^{|a|} \Gamma[(l - k(d_a(j) + 1) - j + 1)/2]$, with $d_a(j)$ equal to the subset specific in-degree of vertex j in a. A multiplicative constant, common to all graphs and k, has been discarded, and $m = p(K + 1)$ yields the minimal training fraction $b = m/n$.

The maximum likelihood estimate $\hat{\Sigma}$ of Σ under the restrictions imposed by $G^{\rightleftharpoons}$ can be computed as follows. Conditional on Σ, the ML estimate of free parameters of $\Pi_1, ..., \Pi_k$ under $G^{\rightleftharpoons}$ is given in Lütkepohl (1993, Sect. 5.2.3) and the ML estimate of Σ under the restrictions in $G^{\rightleftharpoons}$ conditional on $\Pi_1, ..., \Pi_k$ is given in Lauritzen (1996, Proposition 5.9). The ML estimate $\hat{\Sigma}$ is thus obtained by iterating between these two conditional estimators until convergence. In the empirical illustrations in Section 4 convergence was generally reached within a few iterations.

In at least two cases the approximate FML is equal to the exact FML: first, within the subclass of $\mathcal{G}$ treated in Lemma 2 we have $d_c(j) + 1 = |c|$ for all $c \in \mathcal{C}(G^{\sim})$ and the approximate FML is thus equal to the exact FML; second, conditional on the complete graph, the approximate FML of the lag length reduces to the exact FML in Villani (2001a). The following theorem further supports the validity of our approximation.

Theorem 3. The posterior mode estimator of $(G^{\rightleftharpoons}, k)$ based on the approximate FML is weakly consistent.

Proof. See the appendix. ∎

4. Illustrative examples

Example 1. Macroeconomic data

We illustrate the graphical VAR approach on the four-dimensional macroeconomic system of the Danish economy in Johansen (1995). The data consist of 55 quarterly detrended observations from 1974:1 to 1987:3 on log real money, measured by M2, (m), log real GDP (y), the bond rate (b) and bank deposit rate (d). The modulus of the largest eigenvalues of the VAR companion matrix for $k = 2$ are .921, .748 and .748, indicating a stationarity process.

The upper bound for the lag length was set to four and the joint posterior distribution of G and k was computed using the approximate FML. For a fixed lag length there are 262144 possible GCVAR models and 249856 of them are decomposable. The marginal posterior distribution of k is $p(k = 0|X) = .000$, $p(k = 1|X) = .043$, $p(k = 2|X) = .830$, $p(k = 3|X) = .118$ and $p(k = 4|X) = .010$, so the upper bound $k = 4$, does not appear to be restrictive. It is interesting to compare this marginal distribution with the posterior distribution of k conditional on the complete causality graph, which is the model under which the lag length is usually determined. Conditional on the complete graph, $p(k = 1|X) = .559$, $p(k = 2|X) = .441$ and essentially zero probability for other k. A simultaneous analysis of lag length and graph structure thus shifts the probability

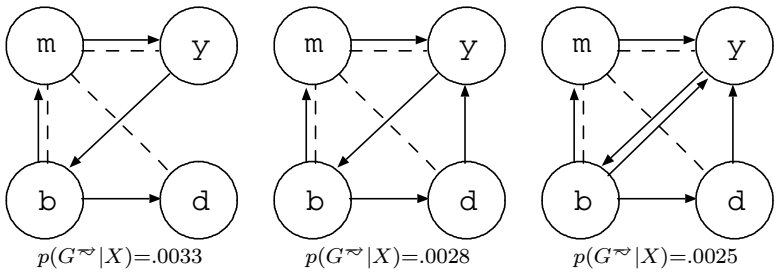

FIGURE 3. Causality graphs with highest posterior probability for the macroeconomic data

Directed edges

	m	y	b	d
m	–	.263	.996	.292
y	.733	–	.322	.631
b	.347	.721	–	.251
d	.396	.245	.987	–

Undirected edges

	m	y	b	d
m	–	.956	.966	.696
y		–	.460	.419
b			–	.481
d				–

TABLE 1. Marginal posterior probabilities of the presence of edges in the macroeconomic data. The directed edges are directed from the variables in the column labels to the variables in the row labels.

mass to larger k compared to the usual analysis conditional on the complete graph. This is of course entirely natural: increasing the lag length under a less than complete graph is not as costly in terms of lost degrees of freedom as under the complete graph where every increase in lag length adds another p^2 parameters to the model. The standard practice of testing zero restrictions, *e.g.* causality restrictions, on the parameters *after* a lag length has been selected is thus likely to be suboptimal.

The most probable causality graphs and their marginal posterior probabilities are displayed in Figure 3. The posterior probabilities of the most probable graphs are not as small as they may seem at first sight when the size of the model space is taken into consideration. A useful benchmark for comparison is the uniform distribution over the set of all causality graphs (for a fixed k) which assigns a probability of roughly $4 \cdot 10^{-6}$ to each causality graph. Another measure of precision in the posterior distribution of $G^{\rightrightarrows}$ is that 5% of the graphs account for approximately 93% of the total probability mass.

Perhaps the most striking feature of Figure 3 is the appearance of the same simple structure of the E_2-graph in all three causality graphs, conveying the message that conditional on a money innovation, all other innovations are independent. The marginal posterior probability of this particular E_2-graph is .131 which should be compared to the benchmark of $1/61 \approx .016$ in the uniform distribution. The second, third and fourth most probable E_2-graphs were all extensions of the simple structure in Figure 3 with exactly one of the undirected edges between y, b and d added to the graph. The posterior probabilities of these graphs were all around .08.

The presence of an edge between any two specific variables is most accurately inferred from the marginal posterior probability of this hypothesis. Table 1 displays these posterior probabilities, both for directed and undirected edges.

	m	y	b	d
m	–	.931	.002	.968
y	.017	–	.135	.079
b	.542	.093	–	.713
d	.322	.882	.000	–

TABLE 2. p-values from Granger non-causality tests conditional on $k = 2$ for the macroeconomic data.

An important issue in macroeconomics is whether or not money has any effect on real variables such as real GDP; see Walsh (1998, Ch. 1) for a review of some empirical evidence and further references. The posterior probability $p(m \to y|X) = .733$ in Table 1 indicates that money probably does matter for real activity. Furthermore, there is much weaker support for the reverse causality from y to m, which is in line with the results of Sims (1972).

It can also be seen from Table 1 that the bond rate is almost certain to be causal for both money and the deposit rate. One way to interpret this finding is that the bond rate is acting as a proxy for expected future inflation. The central bank acts on signals of changes in the inflation rate and their interventions causes movements in both the money stock and the deposit rate.

It is interesting to compare the posterior probabilities of directed edges in Table 1 with classical hypothesis test of non-causality with the absence of a particular directed edge as the null hypothesis (Lütkepohl, 1993). The p-values in Table 2 agree fairly well with the marginal posterior in Table 1. The main difference concerns the relative evidence of the two edges $m \to y$ and $y \to b$, where the p-value of the latter hypothesis is larger than the p-value of the former even though the two hypothesis receive almost the same posterior probability. It should be kept in mind, however, that the hypothesis tests are conditional on $k = 2$ whereas the Bayesian analysis is marginalized with respect to k.

Example 2. Air pollution data.

As a second example, we reconsider here the dependence structure of an air pollution data set investigated earlier in Dahlhaus (2000) using partial correlation graphs. The data involves 4386 daily (4-hour interval) measurements of CO, NO, NO_2, O_3 and the global radiation intensity GRI; for a detailed description of the data, see Dahlhaus (2000). A nonparametric causal analysis of air pollution using 30-minute interval observations can be found in Dahlhaus and Eichler (2001).

Due to the astronomic size of the model space, we investigated whether the lag length could be settled prior to the analysis of the graph structure. The fractional posterior distribution of k (Villani, 2001a) under a uniform prior on $\{0, 1, ..., 10\}$ for k, has its mass completely concentrated on the value 7, which reflects the daily cycle of the observations. The concentration is quite expectable given the large number of observations, and the graph structure is thereby investigated conditional on $k = 7$.

Results of a classical Granger causality tests, are given in Table 3 conditional on $k = 7$. According to these values, only three directed edges can be excluded at a 5%-significance level. However, the battery of tests is subject to the multiple hypothesis testing problem and it is difficult to assess the reasonability of a model where several edges are excluded as a whole. Furthermore, the controversial behavior of p-values for large data sets noticed in the statistical literature, makes their interpretation problematic, see *e.g.* Berger and Sellke (1987).

	CO	NO	NO_2	O_3	GRI
CO	–	.000	.000	.003	.000
NO	.000	–	.038	.073	.000
NO_2	.000	.000	–	.000	.000
O_3	.052	.849	.000	–	.000
GRI	.000	.000	.016	0.000	–

TABLE 3. p-values from Granger non-causality tests conditional on $k = 7$ for the air pollution data.

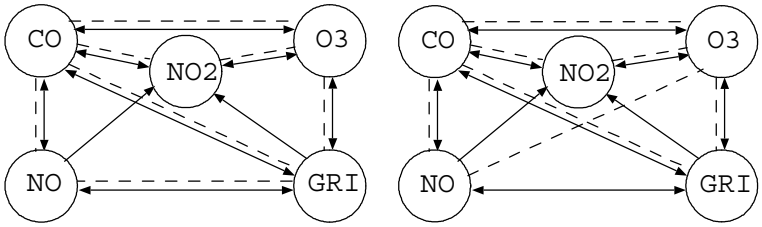

FIGURE 4. The two models with highest posterior probabilities in the class of plausible models for the air pollution data.

For a given lag length, the number of possible causality graph on five time series is 2^{30}, which is currently beyond the capability of commonly used computers. Also, the MCMC methods frequently applied in the more traditional graphical modelling are not expected to provide a practical solution to the current inference problem. Nevertheless, several flexible *heuristic* search algorithms along the lines of Madigan and Raftery (1994) may be used. One possible algorithm will be illustrated here, taking the complete graph as a benchmark for comparison. Define a model to be *plausible* when its approximate FML is higher than that of the complete graph. For each model found to be plausible, the plausibility of submodels with exactly one directed or undirected edge less is investigated, unless they have already been investigated as submodels of some other model during the execution of the algorithm. The algorithm iteratively adds models to the class of plausible models, investigates the submodels of the new models, and terminates when no further plausible submodel can be found.

For the air pollution data, a total of 3744 investigated models were found to be plausible, and the two models with the highest posterior probabilities are presented in Figure 4. To summarize the results of the heuristic model search, the marginal posterior probability of the presence of each edge is also given in Table 4. A rather clear conclusion is that NO and ozone are not directly impacting each other, reflected by the low probability of a directed edge in either direction.

Different conclusions are reached when edges are excluded one at a time. We computed the approximate FMLs of each individual model where exactly one directed or undirected edge is excluded. In Figure 5, we have jointly excluded all edges for which the comparison with the complete graph leads to a larger marginal likelihood for the simpler model. This leads to a considerably simpler graph structure. Notice, that none of the directed edges corresponding to a high p-value appear in the graph in Figure 5. However, several of the absent edges correspond to a very low p-value ($<.0001$), which is well in concordance with the controversial behavior of p-values.

Directed edges

	CO	NO	NO_2	O_3	GRI
CO	–	1.00	1.00	.810	1.00
NO	1.00	–	1.00	.083	1.00
NO_2	1.00	1.00	–	1.00	1.00
O_3	.701	.003	1.00	–	1.00
GRI	.975	.984	.022	1.00	–

Undirected edges

	CO	NO	NO_2	O_3	GRI
CO	–	1.00	1.00	.910	.999
NO		–	.099	.662	.626
NO_2			–	1.00	1.00
O_3				–	1.00
GRI					–

TABLE 4. Marginal posterior probabilities of the presence of edges in the air pollution data. The directed edges are directed from the variables in the column labels to the variables in the row labels.

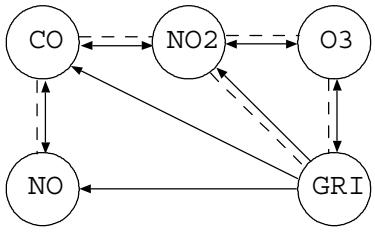

FIGURE 5. Causal graph for the air pollution data based on pairwise exclusion of edges.

5. Concluding remarks

The posterior distribution over $G^{\rightrightarrows}$ and k should be useful for forecasting based on VAR's. The documented poor forecasting performance of VAR models is usually attributed to over-parametrization. One way to combat the resulting erratic parameter estimates and prediction paths was proposed by Litterman (1986) who designed a shrinkage prior on the VAR coefficients which only required modest amounts of subjective inputs from the user. Another way to smooth the VAR predictions is to use Bayesian model averaging (Draper, 1995) to produce a forecast path as a weighted average of the prediction paths under each model $G^{\rightrightarrows}$ and lag length k, where the weights are the posterior probability of the corresponding pair $(G^{\rightrightarrows}, k)$. Measures of prediction uncertainty may be averaged in the same fashion, see Villani (2001b) in the context of cointegration models.

Extensions of graphical models which includes latent variables have recently been proposed in the sphere of cross-sectional analysis, see e.g. the graphical factor analysis models in Giudici and Stanghellini (2002). We are currently working on similar extensions within the fields of time series analysis with particular emphasis on the common trends model for partially non-stationary processes (Stock and Watson, 1988).

Appendix A. Proof of Theorem 3.

To prove the weak consistency of the posterior mode estimator we first establish the asymptotic behavior of the approximate FML. Let $\hat{\Sigma}$ denote the maximum likelihood estimate of Σ under the graph with cliques $\mathcal{C}(G^{\sim})$ and separators $\mathcal{S}(G^{\sim})$. Using that $|\hat{\Sigma}|$ is $\mathrm{O}_p(1)$ and the following identity (Lauritzen, 1996, p. 145)

$$\sum_{c\in\mathcal{C}(G^{\sim})} \log\left|\hat{\Sigma}_{c(k)}\right| - \sum_{s\in\mathcal{S}(G^{\sim})} \log\left|\hat{\Sigma}_{s(k)}\right| = \log|\hat{\Sigma}|.$$

the logarithm of the approximate FML can be written

$$\begin{aligned}
\log m_b(G^{\rightleftharpoons}, k, X) &\propto -\frac{n-m}{2}\left(\sum_{c\in\mathcal{C}(G^{\sim})} \log\left|\hat{\Sigma}_{c(k)}\right| - \sum_{s\in\mathcal{S}(G^{\sim})} \log\left|\hat{\Sigma}_{s(k)}\right|\right) \\
&\quad + \sum_{c\in\mathcal{C}(G^{\sim})} \log\left(\frac{\Gamma^*_{|c|}(n)}{\Gamma^*_{|c|}(m)}\right) - \sum_{s\in\mathcal{S}(G^{\sim})} \log\left(\frac{\Gamma^*_{|s|}(n)}{\Gamma^*_{|s|}(m)}\right) \qquad \text{(A.1)} \\
&= -\frac{n}{2}\log|\hat{\Sigma}| + \sum_{c\in\mathcal{C}(G^{\sim})} \log\Gamma^*_{|c|}(n) - \sum_{s\in\mathcal{S}(G^{\sim})} \log\Gamma^*_{|s|}(n) + \mathrm{O}_p(1),
\end{aligned}$$

where the proportionality sign signals that unimportant additive constants, not depending on either the graph structure or the lag length, have been discarded. Following Villani (2001a), we may use Stirling's formula to approximate $\Gamma^*_{|a|}(h) = \prod_{j=1}^{|a|}\Gamma[(h - k(d_a(j)+1) - j + 1)/2]$ as follows

$$\begin{aligned}
\log\Gamma^*_{|a|}(h) &= \frac{1}{2}\sum_{j=1}^{|a|}[h - k(d_a(j)+1) - j]\log\left(\frac{h - k(d_a(j)+1) - j + 1}{2}\right) \\
&\quad -\frac{1}{2}\sum_{j=1}^{|a|} h - k(d_a(j)+1) - j + 1 + \mathrm{O}(1) \\
&\propto -\frac{1}{2}\sum_{j=1}^{|a|}[k(d_a(j)+1) + j]\log h - \frac{h(1-\log h)|a|}{2} + \mathrm{O}(1). \qquad \text{(A.2)}
\end{aligned}$$

By defining $r(\boldsymbol{\theta}_a)$ as the numbers of unrestricted parameters in the marginal model for subset a, the total number of unrestricted parameters in the model with $(G^{\rightleftharpoons}, k)$ may be written

$$r(\boldsymbol{\theta}) = \sum_{c\in\mathcal{C}(G^{\sim})} r(\boldsymbol{\theta}_c) - \sum_{s\in\mathcal{S}(G^{\sim})} r(\boldsymbol{\theta}_s),$$

since $\sum_{s\in\mathcal{S}(G^{\sim})} r(\boldsymbol{\theta}_s)$ subtracts the number of parameters counted multiple times in $\sum_{c\in\mathcal{C}(G^{\sim})} r(\boldsymbol{\theta}_c)$. By noting that $\sum_{j=1}^{|a|} j$ equals the number of non-redundant elements in Σ_a, and that $k\sum_{j=1}^{|a|}(d_a(j)+1)$ equals the number of predictors used in the marginal model for subset a, we have from (A.1) and (A.2) that

$$\log m_b(G^{\rightleftharpoons}, k, X) \propto -\frac{n}{2}\log|\hat{\Sigma}| - \frac{r(\boldsymbol{\theta})}{2}\log n - \frac{n(1-\log n)}{2}\left(\sum_{c\in\mathcal{C}(G^{\sim})}|c| - \sum_{s\in\mathcal{S}(G^{\sim})}|s|\right) + \mathrm{O}_p(1).$$

Using that $\sum_{c\in\mathcal{C}(G^{\sim})}|c|-\sum_{s\in\mathcal{S}(G^{\sim})}|s|=p$ we obtain the following asymptotic formula for the approximate FML

$$\log m_b(G^{\rightleftharpoons},k,X)\propto-\frac{n}{2}\log|\hat{\Sigma}|-\frac{r(\boldsymbol{\theta})}{2}\log n+\mathrm{O}_p(1) \tag{A.3}$$

Kim (1998) shows that a weakly consistent criterion for model determination in a rather general framework, including the current one, is given by selecting the model which maximizes the expression

$$\log L(X|\hat{\theta},G^{\rightleftharpoons},k)-\log\left(\prod_{l=1}^{r(\boldsymbol{\theta})}h_l(n)\right), \tag{A.4}$$

where $\hat{\theta}$ is the ML estimate of the model parameter vector θ, $h_l(n)$ is the rate of convergence of the maximum likelihood estimate, $\hat{\theta}_l$, of the lth component of $\hat{\theta}$. The models in this paper have $\sqrt{n}$-convergence on all the free parameters under the graph restrictions (see Lütkepohl, 1993, Theorem 5.5 and Lauritzen 1996, formula 5.50), so that $h_l(n)=\sqrt{n}$, for $l=1,...,r(\boldsymbol{\theta})$. Since $\log L(X|\hat{\theta},G^{\rightleftharpoons},k)\propto-\frac{n}{2}\log|\hat{\Sigma}|$, the weakly consistent criterion of Kim (1998) in (A.4) reduces to the asymptotic expression of the approximate FML in (A.3), which in turn shows the weak consistency of the posterior mode estimator based on the approximate FML in Definition 4.

References

[1] Andersson, S. A., Madigan, D. and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.*, **28**, 33-85.

[2] Berger, J. and Sellke, T. (1987). Testing of a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Stat. Assoc.* **82**, 112-139.

[3] Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian theory. Chichester: Wiley.

[4] Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*, McGraw Hill, New York.

[5] Corander, J. (2001). Bayesian graphical model determination using decision theory. To appear in the *J. Multiv. Analysis*.

[6] Corander, J. and Villani, M. (2001). Bayesian assessment of dimensionality in multivariate reduced rank regression. Research report, 2001-2, Stockholm University.

[7] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, **51**, 157-172.

[8] Dahlhaus, R. and Eichler, M. (2001). Causality and graphical models in time series analysis. To appear in: Green, P. J., Hjort, N. L. and Richardson, S. (Eds.): Highly structured stochastic systems. Oxford: Oxford University Press.

[9] Dawid, A.P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.*, **21**, 1272-1317.

[10] Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, 615-633.

[11] Draper, D. (1995). Assessment and propagation of model uncertainty. *J. R. Statist. Soc. B.*, **57**, 45-97.

[12] Eichler, M. (1999). Graphical models in time series analysis. Doctoral thesis, University of Heidelberg, Germany.

[13] Geisser, S. (1965). A Bayes approach for combining correlated estimates. *J. Amer. Stat. Assoc.* **60**, 602-607.

[14] Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, **86**, 785-801.

[15] Giudici, P. and Stanghellini, E. (2002). Bayesian inference for graphical factor analysis models. *Psychometrika, **66**, 577-592.*

[16] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, **37**, 24-36.

[17] Kim, J-Y. (1998). Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models. *Econometrica*, **66**, 359-380.
[18] Lauritzen, S. L. (1996). Graphical models. Oxford: Oxford University Press.
[19] Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions - Five years of experience. *Journal of Business and Economic Statistics*, **4**, 25-38.
[20] Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
[21] Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Stat. Assoc.* **89**, 1535-1546.
[22] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.
[23] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Roy. Statist. Soc.* **B 57**, 99-138.
[24] O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors, *Test*, **6**, 101-118.
[25] Sims, C. A. (1972). Money, income and causality. *American Economic Review*, **62**, 540-552.
[26] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, **48**, 1-48.
[27] Stock, J. H. and Watson, M. W. (1988). Testing for common trends, *J. Amer. Stat. Assoc.,* **83**, 1097-1107.
[28] Villani, M. (2001a). Fractional Bayesian lag length inference in multivariate autoregressive processes. *J. Time Ser. Anal.*, **22**, 67-86.
[29] Villani, M. (2001b). Bayesian prediction with cointegrated vector autoregressions. *Int. J. of Forecasting*, **17**, 585-605.
[30] Walsh, C. E. (1998). *Monetary Theory and Policy*. Cambridge: MIT Press.
[31] Wermuth, N. (1998). Graphical Markov models. In Kotz, S., Read, C. and Banks, D. (Eds.) *Encyclopedia of Statistical Science*. Update Vol. 2. New York: Wiley, 284-300.
[32] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Rolf Nevanlinna Institute, P.O.Box 4, FIN-00014, University of Helsinki, Finland
E-mail address: jukka.corander@rni.helsinki.fi

Department of Statistics, Stockholm University, S-106 91 Stockholm, Sweden
E-mail address: mattias.villani@stat.su.se