

# A BAYESIAN APPROACH TO MODELLING GRAPHICAL VECTOR AUTOREGRESSIONS

JUKKA CORANDER AND MATTIAS VILLANI

ABSTRACT. We introduce a Bayesian approach to model assessment in the class of graphical vector autoregressive (VAR) processes. Due to the very large number of model 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 our approach is illustrated on a four-dimensional macroeconomic system and five-dimensional air pollution data.

KEYWORDS: Fractional Bayes, Granger causality, graphical models, lag length selection, vector autoregression.

## 1. INTRODUCTION

The Gaussian vector autoregression (VAR) can be written as

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

where  $x_t$  is a  $p$ -dimensional vector of time series observations at time  $t$ ,  $\Pi_i$  are  $p \times p$  coefficient matrices determining the dynamics of the system and  $\varepsilon_1, \dots, \varepsilon_n \stackrel{iid}{\sim} N_p(0, \Sigma)$ . 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  $\Pi_1, \dots, \Pi_k$  and  $\Sigma$ . 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 Granger causality tests (Granger, 1969; Sims, 1972).

Our focus here is on Granger 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

---

University of Helsinki (Corander), and Sveriges Riksbank and Stockholm University (Villani). Address for correspondence: Jukka Corander, Department of Mathematics and Statistics, P.O.Box 68, FIN-00014, University of Helsinki, Finland. E-mail: jukka.corander@helsinki.fi. The authors are grateful to Prof. H. Karrasch, Geographisches Institut, for the air pollution data set, and to an anonymous referee whose comments and suggestions significantly improved the original manuscript. Research of the first author has been supported by the funds of University of Helsinki, and 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. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

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 more feasible to consider directed flow, or Granger causality, and it is therefore natural to set up a graph for a time series system with this notion defining the relations between variables.

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 (Dahlhaus, 2000, Dahlhaus and Eichler, 2003, Eichler, 2001, 2002). 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. Recently, Bach and Jordan (2004) introduced a learning algorithm based on a similar spectral representation as in Dahlhaus (2000). However, no attempts to use Bayesian methods seem to have appeared. As inference in graphical models is often 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), Corander (2003), and Tarantola (2004). 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 Granger 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, Corander and Villani, 2004).

The paper is organized as follows. In Section 2 we specify graphical VAR models, and in Section 3 the joint posterior distribution for the lag length and the Granger 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 GRANGER CAUSAL VAR MODELS

Eichler (2001, 2002), and Dahlhaus and Eichler (2003) introduced a general class of Granger causality graphs, including graphical VAR models as a subclass. In the specification of such models, we follow the framework of Dahlhaus and Eichler (2003). Let  $V = \{1, \dots, p\}$  be a set of vertices, representing indices of the stochastic processes in (1.1). In the sequel, we assume that the multivariate process in (1.1) is stationary, and that  $\Sigma$  is positive definite. For an arbitrary subset  $A \subseteq V$ ,  $X_A$  denotes the sub-process

$\{X_A(t), t = 1, 2, \dots\}$  given by the indices in  $A$ , and  $\overline{X}_A(t) = \{X_A(s), s < t\}$  defines the history of  $X_A$  at time  $t$ .

Let  $E_1$  and  $E_2$  be subsets of  $\{V \times V\}$ , denoting the directed and undirected, respectively, edges of a mixed graph  $G = (V, E)$  on  $V$ , where  $E = E_1 \cup E_2$ . Here,  $(a, b) \in E_1$  is interpreted as a directed edge from  $a$  to  $b$ , for all  $a, b \in V$  with  $a \neq b$ . We use  $G^\sim = (V, E_2)$  as a symbol for the undirected part of a mixed graph. Notice, that there are  $2^{\binom{p}{2}}$  undirected and  $2^{3\binom{p}{2}}$  mixed graphs, respectively, for a set of  $p$  vertices. 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, and 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.

The following definition gives the conditions under which a mixed graph  $G = (V, E)$  specifies a *Granger causality graph* for  $X_V$ .

*Definition 1.* A mixed graph  $G = (V, E)$  is a *Granger causality graph* for the process  $X_V$  defined in (1.1), if for all  $a, b \in V$  with  $a \neq b$

- (i)  $(a, b) \notin E_1 \Leftrightarrow X_b(t) \perp \overline{X}_a(t) | \overline{X}_{V \setminus \{a\}}(t)$
  - (ii)  $(a, b) \notin E_2 \Leftrightarrow X_a(t) \perp X_b(t) | \overline{X}_V(t), X_{V \setminus \{a, b\}}(t)$ ,
- where  $\perp$  denotes conditional independence.

The two conditions specify, respectively, that  $X_a$  is not Granger causal for  $X_b$ , and that  $X_a, X_b$  are contemporaneously partially uncorrelated, relative to the process  $X_V$ .

A certain property of undirected graphs, called decomposability, is extremely useful for inferential purposes. An undirected graph is said to be decomposable if it has no chordless cycles of length larger than three. When the part  $G^\sim$  of a mixed graph  $G = (V, E)$  is decomposable, the joint density of  $X_V(t)$  conditional on  $\overline{X}_V(t)$ , factorizes according to (Lauritzen, 1996)

$$(2.1) \quad p(X_V(t) | \overline{X}_V(t)) = \frac{\prod_{C \in \mathcal{C}(G^\sim)} p(X_C(t) | \overline{X}_V(t))}{\prod_{S \in \mathcal{S}(G^\sim)} p(X_S(t) | \overline{X}_V(t))},$$

where  $\mathcal{C}(G^\sim)$  and  $\mathcal{S}(G^\sim)$  are the sets of cliques and separators, respectively, of the graph  $G^\sim$ . The separators of  $G^\sim$  are obtained as intersections of the successive cliques in a so called *perfect* sequence (see, Lauritzen, 1996). Granger causality graphs with decomposable  $G^\sim$  are here termed decomposable Granger causality graphs.

Parametric characterization of a Granger causality graph is provided by the following lemma.

*Lemma 1.* For a Gaussian VAR process  $X_V$  with a Granger causality graph  $G(V, E)$ , the following holds

- (i)  $(a, b) \notin E_1 \Leftrightarrow \Pi_i(a, b) = 0$ , for all  $i = 1, \dots, k$ ,
  - (ii)  $(a, b) \notin E_2 \Leftrightarrow \Omega(a, b) = 0$ ,
- where  $\Pi_i(a, b)$  is the  $(a, b)$ th element of  $\Pi_i$  and  $\Omega = \Sigma^{-1}$ .

*Proof.* Follows directly from Corollary 2.2.1 in Lütkepohl (1993) and Corollary 6.3.4 in Whittaker (1990).  $\square$

A Gaussian VAR process satisfying the conditions in Lemma 1 is said to belong to a graphical Granger causal VAR model, or  $\text{VAR}(G, k)$  for short.

3. BAYESIAN MODEL ASSESSMENT FOR VAR( $G, k$ )

The unknown quantities of a VAR( $G, k$ ) process are: the underlying Granger causality graph,  $G$ , the number of lags,  $k$ , and, conditional on a  $(G, k)$ -pair, the elements in  $\Pi_1, \dots, \Pi_k$  and  $\Sigma$  which are unrestricted under  $G$ . For notational simplicity we use  $\theta$  as a shorthand for the free parameters of a VAR( $G, k$ ) process. Clearly, inference on  $G$  and  $k$  must be settled, before  $\theta$  can be considered.

The main purpose of this paper is to derive the joint posterior distribution of  $(G, k)$  conditional on the observed time series  $\mathbf{X}$ . Let  $K$  be an *a priori* specified upper bound for the value of  $k$ . Using Bayes rule, the joint posterior distribution of  $G$  and  $k$  can be written as

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

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

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

is the marginal likelihood of the observed time series  $\mathbf{X}$ , where  $L(\mathbf{X}|\theta, G, k)$  is the usual likelihood function under model  $(G, k)$  with parameters  $\theta$ , and  $\pi(\theta|G, k)$  is the prior distribution of  $\theta$ . The joint prior of  $k$  and  $G$  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

$$(3.2) \quad \pi(G, k) = \pi(k)/|\mathcal{G}|,$$

where  $\pi(k)$  is some discrete distribution over the integers  $k = 0, 1, \dots, K$ . 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 VAR( $G, k$ ) models. As  $\theta$  varies with  $G$  and  $k$ , elicitation of subjective priors for  $\theta$  can be a difficult and time-consuming task. In many problems one can find a relatively rich class of distributions which both leads to a tractable posterior distribution and at the same is quite easily specified via a few hyperparameters. However, this is not the case for the parametric family considered 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  $\Sigma_A$  to denote the  $|A| \times |A|$  submatrix of  $\Sigma$  formed by the rows and columns of  $\Sigma$  corresponding to the set  $A$ , the prior is of the form

$$(3.3) \quad \pi(\theta|G, k) \propto \frac{\prod_{C \in \mathcal{C}(G^{\sim})} |\Sigma_C|^{-(|C|+1)/2}}{\prod_{S \in \mathcal{S}(G^{\sim})} |\Sigma_S|^{-(|S|+1)/2}}.$$

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  $\Sigma$  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  $G$  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)

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

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

Villani (2001a) showed that the FML approach produced favorable results in inference about  $k$  for VAR( $G, k$ ) models with  $G$  complete, compared to widely used AIC and BIC criteria. 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 specifies the FML for a certain subclass of  $\mathcal{G}$ . In the lemma,  $\mathbf{X}$  denotes the  $n \times p$  matrix of row-stacked observations on the  $p$  time series,  $\mathbf{X}_C$  is the  $n \times |C|$  submatrix of  $\mathbf{X}$  consisting of the columns corresponding to the variables in clique  $C$  and  $\mathbf{X}_{C(-l)}$  is  $\mathbf{X}_C$  lagged  $l$  time periods

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

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

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

$$\begin{aligned} \hat{\Sigma}_C &= n^{-1}(\mathbf{X}_C - \mathbf{Z}_C \hat{\Lambda}_C)'(\mathbf{X}_C - \mathbf{Z}_C \hat{\Lambda}_C) \\ \mathbf{Z}_C &= (\mathbf{X}_{C(-1)}, \dots, \mathbf{X}_{C(-k)}) \\ \hat{\Lambda}_C &= (\mathbf{Z}'_C \mathbf{Z}_C)^{-1} \mathbf{Z}'_C \mathbf{X}_C, \end{aligned}$$

a multiplicative constant, common to all  $G$  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.  $\square$

Unfortunately, the marginal likelihood does not factorize under a general VAR( $G, k$ ) 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 a solution to the model assessment problem within the class  $\mathcal{G}$  with a non-empty separator set, we follow Corander and Villani (2004) and derive an approximation of the FML. A corresponding approximation was shown in Corander and Villani (2004) to perform well in the assessment of dimensionality in reduced rank regression, including cointegrated VAR models.

In the sequel, let the *subset specific in-degree* of vertex  $a$  in  $A \subseteq V$  be defined as the difference  $d_A(a)$  between cardinalities  $|b \in V : (b, a) \in E_1| - |\{b \in V \setminus A : (a, b) \in$

$E_2\} \cap \{(b, a) \in E_1\}$ . That is,  $d_A(a)$  counts the total number of directed edges to  $a$  minus the number of directed edges to  $a$  from vertices  $b$  outside  $A$  such that  $b$  is adjacent to  $a$  in  $G^\sim = (V, E_2)$ . The sole purpose of this definition is to enable a specification of the number of predictors used in a model, such that an eventual overlap in the cliques of  $G^\sim$  can be accounted for in our approximation of FML.

*Definition 2.* Assuming the prior (3.3), the approximate fractional marginal likelihood of a  $p$ -dimensional  $\text{VAR}(G, k)$  is

$$m_b(G, k | \mathbf{X}) = \frac{\prod_{C \in \mathcal{C}(G)} \frac{\Gamma_{|C|}^*(n)}{\Gamma_{|C|}^*(m)} \left| \hat{\Sigma}_C \right|^{-(n-m)/2}}{\prod_{S \in \mathcal{S}(G)} \frac{\Gamma_{|S|}^*(n)}{\Gamma_{|S|}^*(m)} \left| \hat{\Sigma}_S \right|^{-(n-m)/2}},$$

where  $\hat{\Sigma}_A$  is the submatrix of the maximum likelihood estimate  $\hat{\Sigma}$  of  $\Sigma$  under the restrictions given by  $G$ , formed by the rows and columns corresponding to  $A$ , and  $\Gamma_{|A|}^*(l) = \prod_{a=1}^{|A|} \Gamma[(l - k(d_A(a) + 1) - a + 1)/2]$ , with  $d_A(a)$  equal to the subset specific in-degree of vertex  $a \in A$ . It is assumed that the re-labeling of the vertices in any subset  $A$  in  $\Gamma_{|A|}^*(l)$ , preserves the ordering of the original labeling in  $V$ . Since a change in the ordering in  $A$  may result in a change in the value of  $\Gamma_{|A|}^*(l)$ , we have fixed the ordering to avoid ambiguity in the definition. Also, a multiplicative constant, common to all  $G$  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  $\Sigma$  under the restrictions imposed by  $G$  can be computed as follows. Conditional on  $\Sigma$ , the maximum likelihood estimate of free parameters of  $\Pi_1, \dots, \Pi_k$  under  $G$  is given in Lütkepohl (1993, Sect. 5.2.3) and the corresponding estimate of  $\Sigma$  under the restrictions in  $G$  conditional on  $\Pi_1, \dots, \Pi_k$  is given in Lauritzen (1996, Proposition 5.9). The maximum likelihood 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 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  $d_C(c) + 1 = |C|$ , for all  $c \in C, C \in \mathcal{C}(G^\sim)$ , 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  $(G, k)$  based on the approximate FML is weakly consistent.

*Proof.* See the appendix. □

#### 4. ILLUSTRATIVE EXAMPLES

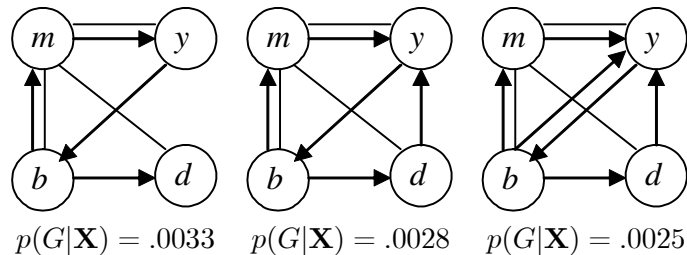
*Example 1.* Macroeconomic data

We illustrate the graphical VAR approach first 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 stationary process (Lütkepohl, 1993).

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  $\text{VAR}(G, k)$  models and 249856 of them are decomposable. The marginal posterior distribution of  $k$  is  $p(k = 0|\mathbf{X}) = .000$ ,  $p(k = 1|\mathbf{X}) = .043$ ,  $p(k = 2|\mathbf{X}) = .830$ ,  $p(k = 3|\mathbf{X}) = .118$  and  $p(k = 4|\mathbf{X}) = .010$ , 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 Granger causality graph, which is the model under which the lag length is usually determined. Conditional on the complete graph,  $p(k = 1|\mathbf{X}) = .559$ ,  $p(k = 2|\mathbf{X}) = .441$  and essentially zero probability for other  $k$ . A simultaneous analysis of lag length and graph structure thus shifts the probability mass to larger  $k$  compared to the usual analysis conditional on the complete graph. This is of course entirely natural, since 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.* Granger causality restrictions, on the parameters *after* a lag length has been selected is thus likely to be suboptimal.

In Theorem 3, it was shown that the approximate FML corresponds asymptotically to the well-known BIC criterion introduced by Schwarz (1978). However, it has also been widely recognized that BIC tends to underestimate model dimension when the sample size is small (Hannan and Quinn, 1979, Villani 2001a, Corander and Villani 2004). Villani (2001a), and Corander and Villani (2004) showed that the FML and approximate FML approach performed favorably in this respect compared to BIC. This feature is mainly due to the nonlinearity of the penalty for model dimension used in the FML approach, which efficiently takes into account the amount of curvature in the likelihood function in the neighborhood of the maximum likelihood estimate. To illustrate the comparison of the two criteria in the current framework, we calculated also the approximate joint posterior distribution of  $G$  and  $k$  based on BIC. From this, the marginal posterior distribution of  $k$  gives the probabilities  $p(k = 1|\mathbf{X}) = .852$ ,  $p(k = 2|\mathbf{X}) = .148$ , and practically zero probabilities for the remaining values. Here, it is again evident that the BIC approximation assigns considerably more posterior weight to simpler models.

The most probable Granger causality graphs and their marginal posterior probabilities are displayed in Figure 1. The posterior probabilities of the most probable graphs are fairly small as might be expected, since the number of available observations is only 55. A useful benchmark for comparison is the uniform distribution over the set of all Granger causality graphs (for a fixed  $k$ ) which assigns a probability of roughly  $4 \cdot 10^{-6}$  to each Granger causality graph. However, the posterior uncertainty reflects the fact that conclusive inference about the overall graph structure cannot be easily made on the basis of this data.



Directed edges					Undirected edges				
	$m$	$y$	$b$	$d$		$m$	$y$	$b$	$d$
$m$	—	.263	.996	.292	$m$	—	.956	.966	.696
$y$	.733	—	.322	.631	$y$		—	.460	.419
$b$	.347	.721	—	.251	$b$			—	.481
$d$	.396	.245	.987	—	$d$				—

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

Directed edges					Undirected edges				
	$m$	$y$	$b$	$d$		$m$	$y$	$b$	$d$
$m$	—	.115	.999	.124	$m$	—	.981	.964	.469
$y$	.255	—	.142	.857	$y$		—	.284	.149
$b$	.310	.491	—	.232	$b$			—	.616
$d$	.322	.233	.977	—	$d$				—

TABLE 2. Marginal posterior probabilities of the presence of edges in the macroeconomic data based on the BIC approximation. The directed edges are directed from the variables in the column labels to the variables in the row labels.

Figure 1. Granger causality graphs with highest posterior probability for macroeconomic data.

Perhaps the most striking feature of Figure 1 is the appearance of the same simple structure of the  $E_2$ -graph in all three Granger 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 1 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. For comparison we calculated also these values based on the BIC approximation (Table 2), and generally, apart from some exceptions, the resulting probability of any edge was considerably lower than the FML based value, which is consistent with earlier findings.

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  $p(m \rightarrow y|X) = .733$  in Table 1 indicates that money probably does matter for real activity. Furthermore, there is

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

TABLE 3.  $p$ -values from Granger non-causality tests conditional on  $k = 2$  for the macroeconomic data.

much weaker support for the reverse Granger causality from  $y$  to  $m$ , which is in line with the results of Sims (1972).

It is interesting to compare the posterior probabilities of directed edges in Table 1 with classical hypothesis test of Granger 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  $m \rightarrow y$  and  $y \rightarrow b$ , 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<sub>2</sub>, O<sub>3</sub> and the global radiation intensity GRI; for a detailed description of the data, see Dahlhaus (2000). A nonparametric Granger causal analysis of air pollution using 30-minute interval observations can be found in Dahlhaus and Eichler (2003).

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  $\{0, 1, \dots, 10\}$  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).

For a given lag length, the number of possible Granger 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

	CO	NO	NO <sub>2</sub>	O <sub>3</sub>	GRI
CO	–	.000	.000	.003	.000
NO	.000	–	.038	.073	.000
NO <sub>2</sub>	.000	.000	–	.000	.000
O <sub>3</sub>	.052	.849	.000	–	.000
GRI	.000	.000	.016	0.000	–

TABLE 4.  $p$ -values from Granger non-causality tests conditional on  $k = 7$  for the air pollution data.

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 2. 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. We also compared here the FML based approximation to the BIC, and contrary to the macroeconomic example, the resulting model probabilities were not qualitatively different for the two approaches. Indeed, this is expected given the large number of observations in the air pollution data set, since the two criteria are asymptotically equivalent.

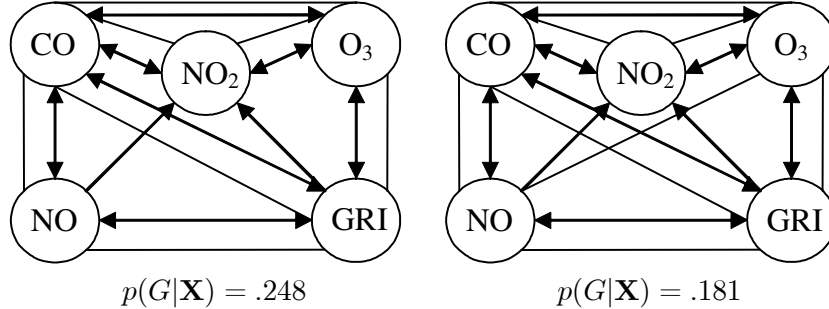


Figure 2. The two models with highest posterior probabilities in the class of plausible models for the air pollution data.

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 3. 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						Undirected edges					
	CO	NO	NO <sub>2</sub>	O <sub>3</sub>	GRI		CO	NO	NO <sub>2</sub>	O <sub>3</sub>	GRI
CO	—	1.00	1.00	.810	1.00	CO	—	1.00	1.00	.910	.999
NO	1.00	—	1.00	.083	1.00	NO		—	.099	.662	.626
NO <sub>2</sub>	1.00	1.00	—	1.00	1.00	NO <sub>2</sub>			—	1.00	1.00
O <sub>3</sub>	.701	.003	1.00	—	1.00	O <sub>3</sub>				—	1.00
GRI	.975	.984	.022	1.00	—	GRI					—

TABLE 5. 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.

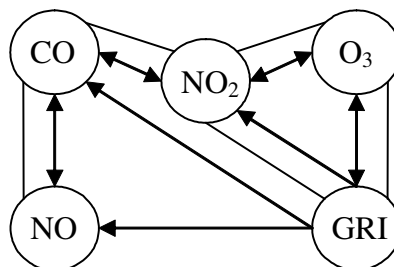


Figure 3. Granger causality graph for the air pollution data based on pairwise exclusion of edges.

## 5. CONCLUDING REMARKS

The posterior distribution over  $G$  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  $G$  and lag length  $k$ , where the weights are the posterior probability of the corresponding pair  $(G, k)$ . 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  $\Sigma$  under the graph with cliques  $\mathcal{C}(G^\sim)$  and separators  $\mathcal{S}(G^\sim)$ . Using that  $|\hat{\Sigma}|$  is  $O_p(1)$  and the following identity (Lauritzen, 1996, p. 145)

$$\sum_{C \in \mathcal{C}(G^\sim)} \log |\hat{\Sigma}_C| - \sum_{S \in \mathcal{S}(G^\sim)} \log |\hat{\Sigma}_S| = \log |\hat{\Sigma}|.$$

the logarithm of the approximate FML can be written

$$\begin{aligned} \log m_b(G, k | \mathbf{X}) &\propto -\frac{n-m}{2} \left( \sum_{C \in \mathcal{C}(G^\sim)} \log |\hat{\Sigma}_C| - \sum_{S \in \mathcal{S}(G^\sim)} \log |\hat{\Sigma}_S| \right) \\ &\quad + \sum_{C \in \mathcal{C}(G^\sim)} \log \left( \frac{\Gamma_{|C|}^*(n)}{\Gamma_{|C|}^*(m)} \right) - \sum_{S \in \mathcal{S}(G^\sim)} \log \left( \frac{\Gamma_{|S|}^*(n)}{\Gamma_{|S|}^*(m)} \right) \quad (\text{A.1}) \\ &= -\frac{n}{2} \log |\hat{\Sigma}| + \sum_{C \in \mathcal{C}(G^\sim)} \log \Gamma_{|C|}^*(n) - \sum_{S \in \mathcal{S}(G^\sim)} \log \Gamma_{|S|}^*(n) + O_p(1), \end{aligned}$$

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  $\Gamma_{|A|}^*(h) = \prod_{a=1}^{|A|} \Gamma[(h - k(d_A(a) + 1) - a + 1)/2]$  as follows

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

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  $(G, k)$  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  $\sum_{a=1}^{|A|} a$  equals the number of non-redundant elements in  $\Sigma_A$ , and that  $k \sum_{a=1}^{|A|} (d_A(a) + 1)$  equals the number of predictors used in the marginal model for subset  $A$ , we have from (A.1) and (A.2) that

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

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

$$(A.3) \quad \log m_b(G, k | \mathbf{X}) \propto -\frac{n}{2} \log |\hat{\Sigma}| - \frac{r(\boldsymbol{\theta})}{2} \log n + O_p(1)$$

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

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

where  $\hat{\boldsymbol{\theta}}$  is the maximum likelihood estimate of the model parameter vector  $\boldsymbol{\theta}$ ,  $h_l(n)$  is the rate of convergence of the maximum likelihood estimate,  $\hat{\theta}_l$ , of the  $l$ th component of  $\hat{\boldsymbol{\theta}}$ . 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  $l = 1, \dots, r(\boldsymbol{\theta})$ . Since  $\log L(\mathbf{X} | \hat{\boldsymbol{\theta}}, G, k) \propto -\frac{n}{2} \log |\hat{\Sigma}|$ , 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 2.

#### REFERENCES

- [1] Bach, F. R. and Jordan, M. I. (2004). Learning graphical models for stationary time series. *IEEE Trans. Signal Process.*, **52**, 2189-2199.
- [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] Corander, J. (2003). Bayesian graphical model determination using decision theory. *J. Multiv. Analysis.*, **85**, 253-266.
- [5] Corander, J. and Villani, M. (2004). Bayesian assessment of dimensionality in reduced rank regression. *Statistica Neerlandica*, **58**, 255-270.
- [6] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, **51**, 157-172.
- [7] Dahlhaus, R. and Eichler, M. (2003). Causality and graphical models in time series analysis. In: Green, P. J., Hjort, N. L. and Richardson, S. (Eds.): *Highly structured stochastic systems*. Oxford: Oxford University Press, 115-137.
- [8] Dawid, A.P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.*, **21**, 1272-1317.
- [9] Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, 615-633.
- [10] Draper, D. (1995). Assessment and propagation of model uncertainty. *J. R. Statist. Soc. B.*, **57**, 45-97.
- [11] Eichler, M. (2001). Graphical modelling of time series, under revision *Scand. J. Stat.*
- [12] Eichler, M. (2002). Granger-causality and path diagrams for multivariate time series, under revision *J. Econ.*
- [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] Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc.*, **B 41**, 190-195.

- [17] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, **37**, 24-36.
- [18] Johansen, S. (1995). Likelihood-based inference in cointegrated vector autoregressive models. New York: Oxford University Press.
- [19] 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.
- [20] Lauritzen, S. L. (1996). Graphical models. Oxford: Oxford University Press.
- [21] Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions - Five years of experience. *Journal of Business and Economic Statistics*, **4**, 25-38.
- [22] Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
- [23] 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.
- [24] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.
- [25] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Roy. Statist. Soc. B* **57**, 99-138.
- [26] O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors, *Test*, **6**, 101-118.
- [27] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461-464.
- [28] Sims, C. A. (1972). Money, income and causality. *American Economic Review*, **62**, 540-552.
- [29] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, **48**, 1-48.
- [30] Stock, J. H. and Watson, M. W. (1988). Testing for common trends, *J. Amer. Stat. Assoc.*, **83**, 1097-1107.
- [31] Tarantola, C. (2004) MCMC model determination for discrete graphical models. *Statistical Modelling*, **4**, 39-61.
- [32] Villani, M. (2001a). Fractional Bayesian lag length inference in multivariate autoregressive processes. *J. Time Ser. Anal.*, **22**, 67-86.
- [33] Villani, M. (2001b). Bayesian prediction with cointegrated vector autoregressions. *Int. J. of Forecasting*, **17**, 585-605.
- [34] Walsh, C. E. (1998). *Monetary Theory and Policy*. Cambridge: MIT Press.
- [35] 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.
- [36] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.