Mach Learn (2009) 75: 217–243
DOI 10.1007/s10994-009-5101-2

Bayesian learning of graphical vector autoregressions with unequal lag-lengths

Pekka Marttinen · Jukka Corander

Received: 25 August 2006 / Revised: 22 December 2008 / Accepted: 27 December 2008 /
Published online: 24 January 2009

Abstract Graphical modelling strategies have been recently discovered as a versatile tool for analyzing multivariate stochastic processes. Vector autoregressive processes can be structurally represented by mixed graphs having both directed and undirected edges between the variables representing process components. To allow for more expressive vector autoregressive structures, we consider models with separate time dynamics for each directed edge and non-decomposable graph topologies for the undirected part of the mixed graph.

Contrary to static graphical models, the number of possible mixed graphs is extremely large even for small systems, and consequently, standard Bayesian computation based on Markov chain Monte Carlo is not in practice a feasible alternative for model learning. To obtain a numerically efficient approach we utilize a recent Bayesian information theoretic criterion for model learning, which has attractive properties when the potential model complexity is large relative to the size of the observed data set. The performance of our method is illustrated by analyzing both simulated and real data sets. Our simulation experiments demonstrate the gains in predictive accuracy which can obtained by considering structural learning of vector autoregressive processes instead of unstructured models. The analysis of the real data also shows that the understanding of the dynamics of a multivariate process can be improved significantly by considering more flexible model classes.

Keywords Bayesian analysis · Granger-causality · Graphical models · Statistical learning · Vector autoregression · Markov chain Monte Carlo · Greedy optimization

Editor: Zoubin Ghahramani.

This work was financially supported by the COMMIT graduate school, the research funds of University of Helsinki, and grant no. 121301 from the Academy of Finland. The authors are grateful to Prof. H. Karrasch, Universität Heidelberg, for the air pollution data set.

P. Marttinen (✉)
University of Helsinki, Helsinki, Finland
e-mail: pekka.marttinen@helsinki.fi

J. Corander
Abo Akademi University Address, Turku, Finland

1 Introduction

Since the late 1970s, extensive multidisciplinary research has crystallized the fundamental versatility of graph representations of multidimensional probability distributions. Such representations, generally referred to as graphical models, intertwine probability theory and graph theory, see Lauritzen (1996) and Jordan (2004). The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model interacting sets of variables, as well as a data structure that lends itself naturally to the design of efficient general-purpose statistical learning algorithms.

As reviewed by Jordan (2004), many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism. Examples of such include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models. The graphical model framework provides a way to view all of these systems as instances of a common underlying formalism. The concept of causality has also been extensively linked to graph theoretical approach through an intervention-based approach, see Pearl (2000) and Spirtes et al. (2000). However, until fairly recently, the core formalism of the graphical models has mostly been anchored to a description of static multivariate probabilistic systems. Early papers considering graphical models in a multivariate time series setting include Lynggaard and Walther (1993) and Brillinger (1996). Since then, numerous refinements and extensions have been considered: Friedman et al. (1998), Stanghellini and Whittaker (1999), Dahlhaus (2000), Reale and Tunnicliffe Wilson (2001, 2002), Bach and Jordan (2004a,2004b), Dahlhaus and Eichler (2003), Fried and Didelez (2003, 2005), Oxley et al. (2004), Moneta and Spirtes (2005), and Carvalho and West (2007). Most of these articles deal with graphs in which the values of the variables at different time stamps are represented by separate nodes in the graph. Our focus is on Granger causality graphs (Eichler 2001) in which each component process is represented by a single node and the interactions between the components are described by a mixed graph, in which directed edges represent Granger causalities and undirected edges represent contemporaneous partial correlations. Similar graphs have been considered in Eichler (2006b, 2007), although in these papers the undirected edges represent contemporaneous covariances between the processes. Strictly speaking, we restrict the analysis to linear Granger causality (Florens and Mouchart 1985), by considering only linear dependencies among the variables.

The undirected dynamic graphical models have recently attracted a vivid interest in functional neuroimaging and patient monitoring in critical care (Gather et al. 2002; Eichler et al. 2003; Salvador et al. 2005; Imhoff and Kuhls 2006; Schelter et al. 2006). The primary reason why the undirected models have gained more attention than the directed models is the mathematical and computational tractability of the structural model learning. Corander and Villani (2006) presented recently a Bayesian method for the learning of directed models. However, their approach is restricted to models which have a decomposable representation in the analogous sense as for the static undirected models (Lauritzen 1996). Also, they did not consider the possibility of allowing the lag lengths to vary over distinct parts of the vector autoregressive system. The restriction to the decomposability of a system has indeed been exploited in the vast majority of all applications of the graphical modelling strategy. The reason for this is typically not an intrinsic connection between the decomposability and some underlying theory of the scientific field in question, but the mathematical and computational tractability of the resulting models. Nevertheless, the restriction to decomposability may lead to unnecessarily complex models having spurious associations between variables as illustrated for static graphical models by Corander (2003).

The Bayesian approach to model learning has in a wide variety of contexts been recognized to yield scientific practices with generally intuitive and appealing characteristics (Bernardo and Smith 1994). In particular, a successful learning of complex statistical models most often necessitates the joint consideration of distinct parts of the models. For instance, such a formulation avoids the multiple hypothesis testing paradox, which often hampers classical data analysis to a large extent. In the context of dynamic graphical models, model learning by minimizing a statistical criterion (e.g. AIC, BIC) has been utilized e.g. by Eichler (2006a). However, especially with mixed graphical presentations, the learning has mostly been based on sequentially testing the exclusion of causal dependencies or partial correlations between variables conditional on the presence of dependencies between all other variables (e.g. Dahlhaus and Eichler 2003; Eichler 2006b). Inevitably, such an approach makes it difficult to assess the reasonability of any model as a whole, and the procedure may lead to distorted dependence structures, as illustrated by Corander and Villani (2006).

The directed models represent a tremendous challenge for model learning as the number of different causal structures increases according to $2^{3\binom{k}{2}}$ as a function of the number of processes (k) in the system. This can be compared to the number of undirected dynamic graphical models $2^{\binom{k}{2}}$, which is the same as in the case of a static system. Despite the apparent challenge in the model learning, the directed dynamic graphical models have considerably more potential to yield meaningful representations of multivariate dynamic systems. As well illustrated by Valdés-Sosa et al. (2005) in the neuroimaging context, the most intriguing questions related to the system structure are statistically represented in terms of causalities, not as partial correlations. Furthermore, the undirected dynamic models have certain mathematical properties which may easily lead to the presence of redundant edges in the estimated graphical representation. As shown theoretically by Dahlhaus and Eichler (2003), these follow from the lacking ability of the undirected models to capture some of the essential features of causality. Mathematically, the redundant edges follow from a generalization of the moralization property of static directed acyclic graphs (Lauritzen 1996).

Although the Bayesian model learning approach can be seen as theoretically preferable to many other existing alternatives, the price of its fundamental coherence is expressed in terms of computational complexity. The recent advances in trans-dimensional Markov chain Monte Carlo (MCMC) methods (Sisson 2005) have largely extended the applicability of the Bayesian approach to model learning. Nevertheless, the current context represents such a numerical challenge that MCMC simulation-based methods are not expected to provide practically applicable solutions. Also, the exploitation of the MCMC approach would necessitate a tedious elicitation of subjective expert knowledge to derive sensible prior distributions for model parameters, as no computationally attractive reference choices exist, in contrast to the static graphical models for discrete data. Hence, the need for unsupervised methods is apparent for dynamic graphical modelling. Therefore, to obtain a numerically feasible learning method, we utilize a recently introduced Bayesian information theoretic criterion for model assessment. This criterion, BEC (Corander and Marttinen 2006), judges the statistical performance of a model in terms of the predictive entropy and can be calculated efficiently even for relatively large systems of variables. To obtain a plausible representation of the graphical representation for a system given an observed multivariate time series, we perform a search in the model space, where the transitions between the putative model structures are governed by comparing the relative performances according to the information theoretic criterion.

The goals of this paper can be summarized as follows: We present a novel learning method for the learning of Granger causality graphs, based on estimation of an underlying graphical vector autoregressive process. The method generalizes the previous approaches by

allowing for non-decomposable undirected part of the graph, and unequal lag-lengths for the underlying process (as opposed to Corander and Villani 2006). We discuss the difficulties related to the learning of such models by standard Bayesian learning based on Markov chain Monte Carlo (MCMC), and propose an algorithm in which the MCMC type calculation is merged with a greedy optimization approach to reach the optimal performance. We investigate the accuracy of the derived criterion, BEC, with some standard alternatives (AIC, BIC, HQ). We perform a simulation study to investigate the significance of allowing unequal lag-lengths in the underlying process, in order to detect the correct Granger causality structure for the process. Finally, we present a real data analysis, which shows that by considering models with unequal lag-lengths, a more expressive representation of the process can be obtained.

This paper is organized as follows. In Sect. 2 we introduce the vector autoregressive processes in the context of directed graphical models. In Sect. 3 the statistical learning process is described. Section 4 provides illustrative examples with both simulated and real data sets, and some concluding remarks are given in the final section.

2 Granger causality graphs for VAR processes

We begin by introducing some graph theoretic concepts and notation. Let $G = (V, E_1, E_2)$ denote a mixed graph (i.e. a graph with both directed and undirected edges), where $V = \{1, \ldots, k\}$ is a set of vertices, $E_1 \subseteq V \times V$ is a set of directed edges and $E_2 \subseteq V \times V$ is a set of undirected edges. Here we allow multiple edges between two nodes: it is possible to have any combination of the following three edges: a directed edge to either direction or an undirected edge. A directed edge (i, i) from node i back into itself is called a loop.

Now, let $G = (V, E_1, E_2)$ be an arbitrary mixed graph. In the current setting, each vertex $i \in V$ represents a univariate stochastic process $X_i = \{X_i(t), t = 1, \ldots, n\}$ from which we have n observations at even intervals. Let $X_V = \{X_i : i \in V\}$ denote the joint multivariate process. For an arbitrary subset $A \subseteq V$, X_A denotes the sub-process $X_A = \{X_i : i \in A\}$, and $\overline{X}_A(t) = \{X_A(s), s < t\}$ denotes the history of X_A at time t. In multivariate time series setting, Granger causality (Granger 1969) offers a tool for the investigation of interactions between the individual processes. These interactions are consistently characterized by Granger causality graphs (Eichler 2001), in which edges encode conditional independences between the process components X_i. More formally, Granger causality graph is defined by two conditions:

Definition 1 A mixed graph $G = (V, E_1, E_2)$ is a *Granger causality graph* for a time series X_V, if, for all $a, b \in V$, $a \neq b$, the following two properties hold for all t:

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

The conditions specify, respectively, that X_a is Granger-noncausal for X_b, and that X_a and X_b are contemporaneously partially uncorrelated. In the original definition, Granger (1969) assumed that the information set contained all the information in the universe, while in practice the interactions are described between a limited number of entities. Thus, a directed edge from a to b does not necessarily mean that a is causing b in the common sense meaning of causality. Such an edge might result from a third variable, not present in the information set, which affects both a and b. Eichler (2008) used a third type of an edge to

describe such spurious causalities. However, here a directed edge from a to b means that the previous values of a contain some information about future values of b, which is not present in any other variable in the information set. Thus, within the given information set, there is a direct relation from a to b. The undirected edges can be understood as interactions which take place on a shorter delay than the interval between two consecutive observations. Further discussion about Granger causality and related issues can be found e.g. in Granger (2001). The complementary intervention-based approach to causality of Spirtes et al. (2000) has been discussed in a dynamic setting by Dash (2005), see also Iwasaki and Simon (1994). To put the Granger causality graphs in the context of dynamic Bayesian network models, it is useful to recall that the model defined above is a chain graph model, where a particular chain component corresponds to $X_V(t)$. Thus, the undirected edges can only exist within each component, whereas any directed edges are always between an element in a component $X_V(t)$ and an element in a subsequent component $X_V(t+r)$, $r > 0$ (i.e. future). As shown in the sequel, by assuming certain types of probabilistic invariances to hold over time, it is possible to retain expressiveness of the model structure while reducing the dimensionality of the learning problem.

Vector autoregressive process with lag-length p, VAR(p), (e.g. Lütkepohl 1993) is defined by

$$x_t = B_1 x_{t-1} + \cdots + B_p x_{t-p} + \epsilon_t, \qquad t = 1, \ldots, n \tag{1}$$

where x_i are $(k \times 1)$ vectors, B_i $(k \times k)$ matrices, $i = 1, \ldots, p$, and x_i, $i = -p+1, \ldots, 0$, are assumed to be available. The vectors ϵ_i are assumed to be i.i.d. $N(0, \Sigma)$. We assume in the sequel that (1) defines a stationary process and that the covariance matrix of the residuals Σ is positive definite. If a VAR process is defined to satisfy the independence conditions present in some Granger causality graph $G = (V, E_1, E_2)$, certain restrictions are imposed on the residual covariance matrix Σ and the coefficient matrices B_i. Such restrictions have been well characterized and are provided by the following lemma.

Lemma 1 *For a VAR(p) process X_V with Granger causality graph $G = (V, E_1, E_2)$, the following hold*

(i) $(a, b) \notin E_1 \Leftrightarrow B_i(b, a) = 0$, *for all* $i = 1, \ldots, p$.
(ii) $(a, b) \notin E_2 \Leftrightarrow \Sigma^{-1}(a, b) = 0$.

Proof See, e.g., Eichler (2001). □

Notice that if $(a, b) \in E_1$, then it follows from (i) of Lemma 1 that $B_i(b, a)$ is non-zero for *some* $i = 1, \ldots, p$. No attempts seem to have been proposed to make this more specific. For example, given that $(a, b) \in E_1$, Corander and Villani (2006) considered learning non-zero parameter values $B_i(b, a)$ for *all* $i = 1, \ldots, p$. In practice, there is usually no reason to believe that lag-lengths for different directed edges would be equal. To improve the VAR learning in this respect, we consider a class of models restricted by two components G and L, where G is a Granger causality graph, and $L \in \mathbb{N}^{k \times k}$ specifies the lag-lengths for the directed edges. More specifically, the entries of L are determined by:

$$(a, b) \notin E_1 \quad \Rightarrow \quad L(a, b) = 0$$

and

$$(a, b) \in E_1 \quad \Rightarrow \quad B_i(b, a) \begin{cases} \neq 0, & \text{for all } i = 1, \ldots, L(a, b), \\ = 0, & \text{if } i > L(a, b). \end{cases}$$

Thus, $L(a, b)$ specifies the lag-length of the directed edge from a to b. In the sequel, we call L as the lag matrix of the process. A VAR process which satisfies the above conditions with respect to a Granger causality graph G and a lag matrix L will be denoted as VAR(G, L). Notice that Definition 1 of Granger causality does not consider self-loops, because they do not affect the Markov properties of the graph. However, in the sequel we will consider explicitly also the lags for the effect of the history of a process component on its own future values, i.e. the presence and absence of the diagonal entries in the coefficient matrices B_i, as determined by the diagonal entries of L. The value $L(a, a)$ will be referred to as the lag-length of the loop $(a, a) \in E_1$.

To introduce some terminology that we use with undirected graphs only, let $\widetilde{G} = (V, E)$ denote the undirected part of a graph G. Let A, B, C and D be arbitrary subsets of V. Let $\widetilde{G}_D = (D, E_D)$ denote a subgraph of $\widetilde{G}$, such that $E_D = \{(i, j) \in E : i, j \in D\}$. The graph $\widetilde{G}_D$ is complete, if $(i, j) \in E_D$ for all $i, j \in D$. A complete subgraph is called a clique, if it is not included in any other complete subgraph. The set C separates A and B if C contains a vertex on every path from A to B. A triple (A, B, C) defines a decomposition of $\widetilde{G}$ if A, B, C are disjoint, $A \cup B \cup C = V$, C separates A from B, and C is a complete subset of V (for details, see Lauritzen 1996, Chap. 2.1). The subgraphs resulting from a decomposition, $\widetilde{G}_{A\cup C}$ and $\widetilde{G}_{B\cup C}$, can be further decomposed until no separating complete subset can be found. A subgraph for which no decomposition can be found is called a prime graph or a prime component. A prime component is maximal if it is not included in any other prime component. Every graph can be uniquely decomposed into maximal prime (mp-) components (Leimer 1993). If all the mp-components are cliques, G is called decomposable (even triangulated, or chordal). Leimer (1993) also describes an algorithm for finding the mp-components of a graph. The usefulness of the concept of decomposition in the learning process of G for a given VAR process X_V is due to the fact that the joint density of $X_V(t)$, conditional on $\overline{X}_V(t)$ factorizes according to

$$p(X_V(t)|\overline{X}_V(t)) = \frac{\prod_{c\in\mathcal{C}(\widetilde{G})} p(X_c(t)|\overline{X}_V(t))}{\prod_{s\in\mathcal{S}(\widetilde{G})} p(X_s(t)|\overline{X}_V(t))}, \tag{2}$$

where $\mathcal{C}(\widetilde{G})$ are the mp-components of $\widetilde{G}$, and $\mathcal{S}(\widetilde{G})$ the corresponding separators (see, e.g. Lauritzen 1996, Chap. 5.2.1). Note that (2) applies even if the mp-components are not cliques. Therefore (2) is useful also when learning non-decomposable graphs, because the mp-components are usually of smaller dimension than the complete graph. However, term $p(X_c(t)|\overline{X}_V(t))$ can be calculated in a closed form in Gaussian setting only if the corresponding component c is complete, otherwise iterative methods must be used.

3 Learning of VAR(G, L)

3.1 Statistical criterion for model plausibility

Our primary statistical learning goal is to be able to identify the model $M_k \in \mathcal{M} = \{M_j : j \in J\}$ in the class of VAR(G, L) models that best describes the relationships between the variables of an observed data set. Each model M_j consists of two primary qualitative components, G, the Granger causality graph, and L, the lag matrix of the VAR-process. An upper bound for the lag-lengths, say p, is assumed to be specified prior to the analysis. We use a model selection criterion, Bayesian Entropy Criterion (BEC) introduced in Corander and Marttinen (2006) to estimate the plausibility of each putative model. According to Bayesian

principles, model choice can be seen as a decision problem where the optimal choice is the model that maximizes the posterior expected utility (Bernardo and Smith 1994)

$$\overline{u}(M_j|\mathbf{x}) = \int_{\Theta_j} u(M_j, \theta_j)\pi(\theta_j|\mathbf{x})d\theta_j, \tag{3}$$

where θ_j comprises the model parameters, i.e. the coefficient matrices B_i, $i = 1, \ldots, p$, and the residual covariance matrix Σ of the VAR process. Here $u(M_j, \theta_j)$ is the utility of model M_j given that θ_j is the true parameter value, and $\pi(\theta_j|\mathbf{x})$ is the posterior distribution of the model parameters given current observations $\mathbf{x}$. Our utility function u uses a logarithmic score (see, e.g. Bernardo and Smith 1994) to measure the expected performance in predicting a future data set $\mathbf{y}$ of a similar structure, and is defined by

$$u(M_j, \theta_j) = \int_{\mathcal{X}} p_j(\mathbf{y}|\theta_j) \log p_j(\mathbf{y}|\theta_j)d\mathbf{y}, \tag{4}$$

where $p_j(\cdot|\theta_j)$ is the predictive probability distribution from model M_j with parameter values θ_j. The considered future data set $\mathbf{y}$ is from the same model, and of the same size as the current data set $\mathbf{x}$.

The characteristic feature of the BEC criterion is that the expectation in (4) is taken with respect to the predictive distribution of model M_j. This has certain advantages in the current setting, e.g. compared to the approach of Corander and Villani (2006), where the approximate Bayesian model learning criterion required training of the prior distributions with respect to the most complex model considered. This would typically correspond to the complete model, having the complete graph G and the largest possible lag-length for each edge, easily leading to a situation where the number of observations would not be sufficient for the estimation of the parameters of the model. As (3) allows us to compute the expected utility of every model per se, this approach avoids the fitting of overly complex models similarly to the common asymptotic model selection criteria such as AIC, BIC and HQ, from Akaike (1969), Schwarz (1978), and Hannan and Quinn (1979), respectively. However, unlike these asymptotic criteria with a linear penalty term with respect to the model complexity for a fixed (observed) sample size, BEC takes the curvature in log-likelihood into account, and behaves non-linearly with respect to the increasing model complexity. This makes BEC an attractive choice when the number of observations is small relative to the complexity of the putative models, as will be illustrated in Sect. 4.

Thus far no prior knowledge about the actual model structure M_j has been incorporated to the model score. This can be done by subtracting a penalty term c_j reflecting model complexity from (3) (Bernardo 1999; Corander and Marttinen 2006). The resulting BEC criterion equals

$$BEC(M_j) = \int_{\Theta_j} u(M_j, \theta_j)\pi(\theta_j|\mathbf{x})d\theta_j - c_j. \tag{5}$$

We use reference priors for Σ and B_i, $i = 1, \ldots, p$, and show in the Appendix that BEC criterion for model M_j can be written as

$$BEC(M_j) = \log p_j(\mathbf{x}|\widehat{\theta}_j) + f(M_j) - c_j, \tag{6}$$

where $f(M_j)$ can be calculated analytically. The first term in sum (6) is the maximized log-likelihood under model M_j. It will be shown in the Appendix that, as n increases, the

difference in the f terms in (6) between two models M_j and M_l converges to a constant value:

$$f(M_j) - f(M_l) \overset{n\to\infty}{\to} \frac{d_l - d_j}{2},$$

where d_l and d_j are the parametric dimensionalities of the two models. Thus, different choices of the penalty term c_j make BEC asymptotically equivalent to other model selection criteria (e.g. AIC, BIC, HQ). We set

$$c_j = d_j \log\log n, \tag{7}$$

which is in VAR framework the slowest increasing asymptotically consistent penalty (see Hannan and Quinn 1979), corresponding to vague prior information about model structure. With this choice, the BEC criterion (6) can be seen as analogous to the HQ criterion, however, with a correction term $f(M_j)$ to improve small sample properties.

Interpretation of the relative plausibilities of different models can be obtained by defining a probability distribution from relative utilities according to

$$q(M_j) = \frac{\exp(BEC(M_j))}{\sum_{M\in\mathcal{M}} \exp(BEC(M))}. \tag{8}$$

It follows from the statistical consistency of the HQ criterion that, if one of the models in $\mathcal{M}$ is the true generating model, its relative utility (8) will converge to unity as the number of observations increases. The relative utilities can be interpreted as asymptotic approximations to posterior probabilities of models, analogously to similarly normalized versions of statistically consistent criteria, such as BIC and HQ.

3.2 Algorithms

Here we describe the algorithm used to search the space of models (G, L) for a VAR(G, L). (The ml-estimation of the parameters, required to calculate BEC (6), is described in the Appendix). A greedy search using steepest descend is a computationally attractive choice in the setting of graphical models (Heckerman et al. 1995), where the vast model space makes standard exact search strategies (e.g. Back-Tracking, Cormen et al. 2001) infeasible. Other methods have been proposed as well, such as Monte Carlo techniques (see, e.g. Janzura and Nielsen 2006) and exact computation under structural model restrictions (Koivisto and Sood 2004). However, the increase in the degree of model complexity induced by the dynamics as compared to regular static graphical models (typically for discrete-valued nodes), makes the model learning even less tractable. A comprehensive collection of various model reduction methods, which are commonly used in econometrics for learning the dynamic part of a VAR, can be found in a comparison study by Brüggemann et al. (2002), see also Winker and Maringer (2004) and Ozcicek and McMillin (1999). Here, we propose an algorithm, which combines a greedy approach for efficiency and MCMC type calculation for consistency.

Let (G, L) be a model describing the structure of a VAR(G, L) process. We will use $(G, L)_{(i,j)=l}$ to denote the model which equals otherwise (G, L), except that the lag-length of directed edge (i, j) is set to l. Now, (8) can be utilized to define a distribution for the lag-length of a directed edge, conditional on other edges, as

$$q(L(i, j) = l|(G, L)) = \frac{q((G, L)_{(i,j)=l})}{\sum_{r=0}^{p} q((G, L)_{(i,j)=r})}. \tag{9}$$

Notice that in $(G, L)_{(i,j)=0}$ the corresponding directed edge vanishes from the model. Similarly, let $(G, L)_{\{i,j\}=1}$ denote the model which is obtained from (G, L) by adding an undirected edge between i and j, and $(G, L)_{\{i,j\}=0}$ denote the model obtained by removing the edge. The conditional probability of the undirected edge is given by

$$q((i, j) \in E_2|(G, L)) = \frac{q((G, L)_{\{i,j\}=1})}{q((G, L)_{\{i,j\}=0}) + q((G, L)_{\{i,j\}=1})}. \tag{10}$$

The above two conditional distributions can be interpreted as approximate conditional posterior distributions over the subclass of models where the condition holds. Our search algorithm proceeds by updating edges one by one, using the distributions (9) and (10). The process of updating the status of all edges once, is referred to as one iteration in the sequel. Such iterations can be performed either in a greedy or a stochastic manner. In a greedy iteration, each edge is updated to maximize the BEC value, whereas in a stochastic iteration the values for the edges are drawn from the distributions (9) and (10). Notice that one stochastic iteration corresponds to one iteration in a standard Gibbs sampling MCMC simulation (see e.g. Robert and Casella 2005) from the distribution (8). In our MATLAB implementation, the algorithm proceeds by alternating between the stochastic and greedy iterations. In practice this strategy leads to convergence quite rapidly, and in the performed simulation experiments (see Sect. 4), no more than two or three stochastic iterations were usually required to reach a state where the algorithm no longer identifies new models in addition to those already associated with high BEC values.

The combination of stochastic and greedy steps provides an important balance for the search algorithm. Namely, if only stochastic steps were exploited in the search, the identification of a model with a very high ranking with respect to BEC would become extremely unlikely in practice. This is due to the fact that, when the search is close to a (locally) optimal model, the number of possible updates worsening the model is considerably larger than the number of updates that would improve the model. Therefore, it is unlikely that all the edges would simultaneously be associated with the optimal values, if these were chosen stochastically. This behavior was in fact clearly present in the analyses we performed with both real and simulated data. On the other hand, using only greedy iterations would most likely lead to a model which is only locally optimal. Indeed, the eventual discovery of the true global optimum is theoretically guaranteed by the stochastic iterations, which allow the search to escape from local maxima. Notice that there is a strictly non-zero probability that, starting from *any* state, after one stochastic iteration the search is in the globally optimal state. It follows that, as the number of performed stochastic iterations increases, the probability of not visiting the true global optimum converges to zero.

Here we consider briefly the time complexity of the presented method. An exact derivation of the time complexity is straightforward and is omitted. Instead, we make the following remarks: (1) Factors affecting the time complexity include: dimension of the process (k), the lag-lengths of the directed edges, the number of data points n, the upper bound for the lag-length (p), and the size and the structure of the undirected part of the graph. (2) Some steps in the ml-estimation of the model parameters (see Appendix) require quadratic or cubic operations, and consequently the algorithm can not be expected to be easily scalable for modeling situations with arbitrarily increasing complexity. (3) The sizes of the mp-components of the undirected part of the graph determine largely the final complexity of the algorithm, especially if the components are non-complete, and consequently require the use of the iterative methods for the ml-estimation of the corresponding parameters in the covariance matrix. The observed execution times as well as the practical limits for the applicability of the presented algorithm are discussed in the end of Sect. 4.2.

For the purposes of the simulations in the next section, we need also such versions of the described algorithm, which can be used for the learning of the structure of either VAR(G, p^*) or VAR($\widetilde{G}$), where VAR(G, p^*) refers to a process in which all lags are assumed equal (p^*), while VAR($\widetilde{G}$) denotes a process, in which only the covariance matrix is restricted by the undirected graph $\widetilde{G}$, while the dynamic part, i.e. the coefficient matrices, is unconstrained. The latter class of models can be interpreted as unstructured VAR models, where the causal effects may be absent or present for any particular lag length, given the presence of least a single non-zero element among $B_i(b,a)$, for $i = 1, \ldots, p$, which would imply the existence of the directed edge (a, b) in the Granger causality graph. (VAR($\widetilde{G}$) can be considered as a DBN with contemporaneous dependencies represented by undirected edges.)

The learning algorithms of VAR(G, p^*) and VAR($\widetilde{G}$) are based on straightforward modifications of the algorithm presented for the learning of a VAR(G, L) process. Although the model space of VAR($\widetilde{G}$) is vast as compared to VAR(G, L), one iteration of the search algorithm is of the same order of complexity as with VAR(G, L). This is due to the fact that updating the presence or absence of p parameters corresponding to one directed edge in the unconstrained process requires the calculation of p bivariate conditional distributions based on (8), while the update of a lag value for one edge in VAR(G, L) requires the calculation of one p-variate distribution. Both the operations require $O(p)$ evaluations of BEC for different models. Further details of these algorithms are omitted.

4 Examples

4.1 Synthetic data

To investigate the accuracy of our method under different conditions, we performed a simulation study with five different types of underlying graphs. For each graph type, we randomly created a set of 20 graphs which were analyzed with the presented algorithm. The different graph types that were considered include: (1) "basic" graph, (2) sparse graph, (3) graph with directed edges only, (4) graph with undirected edges only, and (5) graph with non-decomposable undirected part. The graph types were specified by five parameters: the dimension of a process (k), the maximum lag-length for any directed edge ($p_{\max}$), and probabilities of a directed edge from one node to another (p_1), a loop (p_2), and an undirected edge (p_3). For example, for the "basic" graph type, the above parameters were set to: $k = 5$, $p_{\max} = 5$, $p_1 = 0.3$, $p_2 = 0.7$, and $p_3 = 0.5$. The graphs for each setup were randomly created using the specified parameters. For directed edges present in a graph, the lag-lengths were uniformly drawn from $1, \ldots, p_{\max}$.

The coefficient matrices B_i ($i = 1, \ldots, p_{\max}$) of the VAR processes corresponding to the simulated graphs were drawn from distributions:

$$B_i(a,b) \sim \begin{cases} U(-0.5, 0.5), & \text{if } a = b \text{ and } i \le L(b,a), \\ U(-1, 1), & \text{if } a \ne b \text{ and } i \le L(b,a), \\ 0, & \text{if } i > L(b,a). \end{cases}$$

Only the coefficient matrices satisfying the stability condition (Lütkepohl 1993, formula 2.1.12) were accepted for further analysis. The covariance matrices were simulated with the MATLAB function *sprandsym*, which can be used to generate symmetric positive definite matrices. Iterative proportional fitting was used to transform the inverse covariance

matrices to satisfy the restrictions imposed by the undirected parts of the graphs. Before accepting the generated covariance matrix we checked that the elements in the precision matrix, corresponding to partially contemporaneously correlated variables, were non-zero also in practice. Those non-zero elements of the precision matrix which were less than 0.05 were randomly assigned new values from the distribution $U(0.07, 0.17)$.

Data sets of sizes 200, 1000, and 5000 observations were generated for each process, using the strategy described in Lütkepohl (1993), Appendix D.1. The estimation algorithm was then run with each data set. The search was performed using both the stochastic and the greedy search steps, such that the search was started with three stochastic steps, after which the greedy step was repeated until no improvement occurred. These steps were repeated twice with the data sets having dimension 5, and three times with data sets of higher dimension. The results along with different parameter values are collectively presented in Table 1, and they are based on the 20 simulated graphs of each graph type. To measure the differences between lag matrices we use a norm for matrix $T = (t_{ij})_{i,j=1,\ldots,k}$ defined as:

$$\|T\| = \sum_{i=1}^{k} \sum_{j=1}^{k} |t_{ij}|.$$

Because element (i, j) of the lag matrix specifies the number of parameters in the coefficient matrices which are used to describe the directed edge from i to j, it follows that the norm of the lag matrix $\|L\|$ corresponds to the number of non-zero parameters in the coefficient matrices.

The values in Table 1 show that our method is capable of inferring the correct graph structure well, both the undirected edges, and the directed edges with the lag-lengths. Also, the estimates are closer to the true underlying graphs, when the number of observations increases, as expected from the theoretical perspective. The accuracy of the estimation procedure decreases slightly with an increasing dimension of the process, as can be seen by comparing the graph types with dimension 7 and 10 to those with dimension 5. This reflects the increased complexity, caused by the much larger number of possible models in higher dimensions. However, with an adequate number of observations, the estimated graphs were more accurate also in the higher dimensional situations. It can be seen that the directed edges are inferred with a relatively high accuracy even with the smallest data sets, while the inference of the undirected part seems to require more observations to yield a correct identification of most of the edges. The complete exclusion of either directed or undirected edges does not seem to create any bias, but the estimates of the remaining edges are as good as with mixed graphs of similar complexity.

To illustrate the benefits of using BEC, we investigated the small sample properties of other commonly used criteria, AIC, BIC, and HQ, and compared these to the results obtained by the BEC criterion (6). We also considered various forms of the penalty term c_j in (6), making BEC asymptotically equivalent in turn to each of the mentioned criteria. In total 40 data sets of 50 observations were generated by using the above described simulation setup with the basic graph type. The data sets were analyzed with the described algorithm, changing only the used criterion. The results of these analyses are presented in Table 2. The results show that, regardless of the criterion of choice (AIC, BIC, HQ), the BEC criterion with the corresponding penalty term c_j is able to infer the underlying model structure more accurately than its asymptotic counterpart.

Finally, we investigated the estimation of the Granger causality graph under a range of different constraints on the dynamic part (the coefficient matrices) of a process, when the true underlying graph had unequal lag-lengths. In particular, we considered three different

Table 1 Results with synthetic data. Each row shows average results from 20 graphs. The meanings of the columns are: k: the dimension of the process. $p_{\max}$: the upper bound for the length of the memory of any single process. p_1: the probability of a directed edge from one node to another. p_2: the probability of a loop. p_3: the probability of an undirected edge. For the last setup (*) only the graphs having non-decomposable undirected parts were accepted for the analysis. #data: the number of observations in the data set. $\|L\|$: the average norm of the lag matrices in the simulated graphs. $\|L - \widetilde{L}\|$: the average distance between the true and the estimated lag matrices. Standard deviations are shown in the parentheses. $|E_2|$: the average number of undirected edges in the simulated graphs. Δ_{E_2}: the average number of differing undirected edges between the true and the estimated graphs. Standard deviations are shown in the parentheses

Setup						Results			
k	$p_{\max}$	p_1	p_2	p_3	#data	$\|L\|$	$\|L - \widetilde{L}\|$	$\|E_2\|$	Δ_{E_2}
5	5	0.3	0.7	0.5	200	28.9	4.1(3.2)	5.0	3.6(1.5)
					1000		1.6(1.3)		1.4(1.3)
					5000		1.0(0.9)		0.3(0.6)
10	3	0.05	0.3	0.15	200	15.4	9.1(3.6)	8.1	8.6(3.0)
					1000		4.8(2.6)		4.1(1.6)
					5000		2.9(1.8)		1.3(1.0)
5	5	0.3	0.7	0	200	26.6	3.6(2.0)	0.0	0.4(0.5)
					1000		1.7(1.4)		0.2(0.4)
					5000		0.8(0.7)		0.2(0.4)
5	5	0	0	0.5	200	0.0	1.7(1.4)	4.7	3.5(1.6)
					1000		1.3(1.3)		1.4(1.1)
					5000		0.8(0.9)		0.2(0.4)
7	4	0.2	0.7	0.5*	200	33.0	7.3(3.2)	11.3	8.6(3.0)
					1000		3.0(1.8)		4.0(1.4)
					5000		1.8(1.5)		0.6(0.8)

Table 2 The results of the comparison of small sample behavior between various model choice criteria. The simulation setup corresponding to row 1 in Table 1 was repeated for 40 data sets of 50 observations. $\|L - \widetilde{L}\|$ is the average distance between the true and the estimated lag matrices. Δ_{E_2} is the average number of differing undirected edges between the true and the estimated graphs. BEC_{AIC}, BEC_{BIC} and BEC_{HQ} are variations of BEC, in which the penalty term c_j in (5) is selected to make the criterion asymptotically equivalent to AIC, BIC or HQ, correspondingly. Thus, BEC_{HQ} corresponds to the criterion referred to simply as BEC elsewhere in the paper

Criterion	$\|L - \widetilde{L}\|$	Δ_{E_2}
AIC	80.1	5.1
BEC_{AIC}	47.8	4.7
BIC	16.5	3.9
BEC_{BIC}	12.4	4.3
HQ	46.1	4.8
BEC_{HQ}	17.2	4.1

Table 3 Comparison of the accuracy of learning Granger causality graphs, when the learning is based on VAR processes with differing structural constraints on the dynamic part of the process. In VAR(G, L) the coefficient matrices are constrained by lag matrix L. In VAR(G, p^*) all lags are assumed to be equal. VAR($\widetilde{G}$) corresponds to a VAR with an unconstrained lag-structure. The following parameter values (see Table 1) were used to simulate the generating graphs for the data sets: $K = 5$, $p_1 = 0.6$, $p_2 = 0.9$, $p_3 = 0.5$, and #data $= 300$. The lag-lengths for the edges were generated on the first row from $U(1, \ldots, 5)$ and on the second row from $U(1, 2)$. However, on the second row one of the lags was drawn from $U(4, \ldots, 8)$, and one from $U(18, \ldots, 20)$. The results are based on average values for 20 data sets. Δ_{E_1} and Δ_{E_2} denote the numbers of differing directed and undirected edges between the estimated and the true graph. MSE (mean squared error) measures the predictive performance of the estimated graph

	VAR(G, L)			VAR($\widetilde{G}$)			VAR(G, p^*)		
Lag-length	Δ_{E_1}	Δ_{E_2}	MSE	Δ_{E_1}	Δ_{E_2}	MSE	Δ_{E_1}	Δ_{E_2}	MSE
$U(1, \ldots, 5)$	1.4	2.7	3.7	5.7	2.9	4.0	1.7	2.9	3.9
Mixture	1.5	2.9	3.7	5.4	2.8	4.1	3.5	3.5	5.3

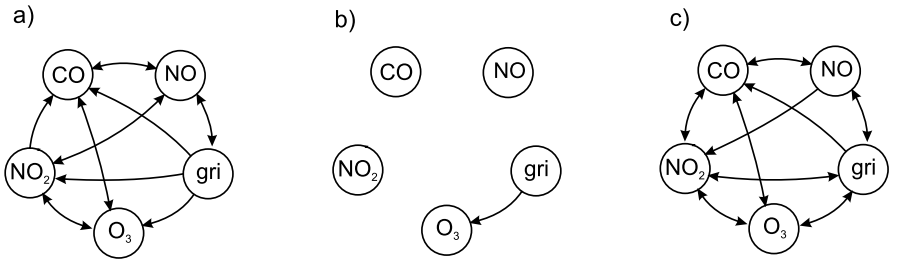
approaches: (1) the traditional approach, in which all lag-lengths are assumed equal, i.e. VAR(G, p^*), (2) the presented approach, where each directed edge has its own lag-length, VAR(G, L), and (3) VAR($\widetilde{G}$), in which the dynamic part of the process is unconstrained. We investigated the estimation of Granger causality graphs in two different underlying lag-length settings: in the first setting, the lag-lengths were drawn from $U(1, \ldots, 5)$. In the second setting the lag-lengths were drawn from $U(1, 2)$, except that one lag-length was drawn from $U(4, \ldots, 8)$ and one from $U(18, \ldots, 20)$. Model structures restricted by upper bound $p = 23$ were considered by the search algorithms. Both settings were repeated 20 times, and the average results are shown in Table 3. The results show that, in both the settings, the estimate of the Granger causality graph based on the unrestricted VAR($\widetilde{G}$) was the most inaccurate. The reason for this is that VAR($\widetilde{G}$) allows for the inclusion of individual parameters in the coefficient matrices, which increases the model complexity only a little and thus leads to a light penalty by the model selection criterion. Consequently, the corresponding causality graphs have numerous redundant directed edges (related issues have been discussed also in Eichler 2001). In contrast, an inclusion of a directed edge in VAR(G, L) or VAR(G, p^*) in general adds many parameters to the model and is therefore penalized more heavily, leading to fewer false associations between the variables. When the lag-lengths were simulated from $U(1, \ldots, 5)$, the assumption that all lag-lengths are equal did not seemingly reduce the accuracy of estimating the Granger causality graph. On the other hand, when some directed edges had clearly larger lag-lengths than others (the second setting), the estimated overall lag-length in VAR(G, p^*) was in general estimated to be somewhere in between the extreme values. Also, the estimated Granger causality graph often did not contain all the edges with shorter lags present in the true graph, because inclusion of such edges would have introduced many redundant parameters, in addition to the actual ones. The learning based on VAR(G, L), allowing for unequal lags, was not corrupted by such behavior. When we used the learned models with ml-parameter values to predict yet another 10 observations from the true underlying models, the uncertainty related to the redundant parameters in VAR(G, p^*) model is clearly visible as a lowered predictive ability. Also the unrestricted VAR($\widetilde{G}$) models had clearly better predictive abilities than the VAR(G, p^*) models, even if the corresponding Granger causality graphs were contaminated by spurious edges. The importance of allowing unequal lag-lengths for VAR models has been discussed also e.g. in Gredenhoff and Karlsson (1999).

4.2 Air pollution data

As a real-world example we analyzed air pollution data previously examined by Dahlhaus (2000), Dahlhaus and Eichler (2003) and Corander and Villani (2006). Dahlhaus (2000) and Dahlhaus and Eichler (2003) used partial correlation graphs in their analysis and Corander and Villani (2006) restricted their fractional Bayesian analysis to cover the Granger causality graphs with decomposable undirected parts of the graph and a common lag-length. We searched for the best model among all graphs with directed and undirected edges, allowing for unequal lag-lengths. The data has been collected by half-hourly measurements, and consists of 48216 observations, thus covering a period of less than three years. The recorded variables include CO, NO, NO_2, O_3 and global radiation intensity (GRI). CO and NO are created mostly by traffic and industry, whereas NO_2 and O_3 are created by different chemical processes in the atmosphere. The global radiation plays a role in these processes, being especially essential for the birth of ozone. For further details of the interactions between the variables, the reader is referred to Seinfeld (1986, Chap. 4). As in Dahlhaus (2000), we performed the analysis with both the original (raw) data and a trend-corrected (residual) data, which was obtained by subtracting a local average daily cycle from the observations. For the original data, only every 8th observation (4 hour interval) was considered in the analysis, as in Dahlhaus (2000) and Corander and Villani (2006). For the trend-corrected data analysis we used all the data points, as in Dahlhaus and Eichler (2003). The detailed results are shown for the original data only. However, differences and similarities between the raw and residual data analyses will be discussed afterwards.

Before the analysis an upper bound p of the lag-lengths is required to be specified. Here we used a lag-length equal to 60 as the upper bound for the loops and 20 for the rest of the directed edges (upper bound 60 corresponds to a period of $60*4$ hours i.e. 10 days). An upper bound equal to 80 was also tested for all the edges, but this did not lead to any change in the results. We ran the search algorithm independently five times, by repeating on each of these twice a succession of 4 stochastic and greedy iterations. The optimal graphs found in the five replicates of the search all had the same undirected part, while the directed parts of graphs varied to some extent over the runs. In two of the replicates, the graph corresponding to the overall optimum over all runs was identified, while in the three others the lag-lengths of some directed edges were slightly different from the optimum. The maximum difference $\|L_1 - L_2\|$ of a lag matrix L_2 found in any single replicate to the overall optimal lag matrix L_1 was 9. This is small as compared to the norm of the optimal lag matrix $\|L_1\|$, which equals 253. Thus, from the practical perspective the algorithm converged to almost the same model in all five replicates. All the graphs visited during the five runs of the algorithm were combined to calculate the relative utilities (8), which can also be interpreted as approximate posterior probabilities as discussed earlier. Due to the vast model space, no single model was clearly preferred over others. The highest value of the approximate probability (8) of any single graph was 0.128. The directed part of this graph is presented in Fig. 1a and the undirected part is presented in Fig. 2a. The marginal approximate probabilities for the undirected and directed edges are given in Tables 4 and 5.

Another interesting way to investigate the data is to calculate the approximate posterior distributions of the lag-lengths of the directed edges, conditional on the graph with the highest BEC value. This offers the advantage that we can investigate the relative plausibility of different lag-lengths over the whole range of possible values, whereas simply calculating the approximate marginal distribution based on the states visited during the search gives zero probability to most of the lag-lengths, because the corresponding states are never visited in practice. Plots of these conditional probabilities are shown in Fig. 3.



	CO	NO	NO_2	O_3	gri
CO	43	1	0	2	0
NO	1	43	1	0	2
NO_2	2	2	42	2	0
O_3	1	0	1	43	0
gri	7	5	5	8	42

	CO	NO	NO_2	O_3	gri
CO	43	0	0	0	0
NO	0	43	0	0	0
NO_2	0	0	43	0	0
O_3	0	0	0	43	0
gri	0	0	0	43	43

	CO	NO	NO_2	O_3	gri
CO	7	7	7	7	0
NO	7	7	7	0	7
NO_2	7	0	7	7	7
O_3	7	0	7	7	7
gri	7	7	7	7	7

Fig. 1 The directed parts of Granger causality graphs for the air pollution data (loops are not shown). The corresponding lag matrix is shown below each graph. (**a**) The directed part of the graph with the highest BEC value. (**b**) The directed part of the graph with the highest BEC value, such that each edge has the same lag length. (**c**) The directed part of the graph with the highest BEC-value, when each edge is constrained to have lag length equal to 7 (a local optimum)

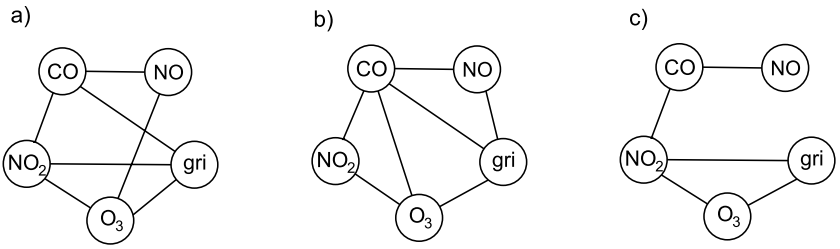


Fig. 2 The undirected parts of Granger causality graphs for the air pollution data. (**a**) The undirected part of the graph with the highest BEC value. (**b**) The undirected part of the graph inferred in Corander and Villani (2006). (**c**) The undirected part of the graph from Corander and Villani (2006), when the edges were inferred using a pairwise testing procedure

Table 4 The marginal probabilities of the undirected edges of the Granger causality graph for the air pollution data

	NO	NO_2	O_3	GRI
CO	1.00	1.00	0.18	1.00
NO		0.01	1.00	0.01
NO_2			1.00	1.00
O_3				1.00

By examining Fig. 1a we can divide the directed edges which are present in the graph with the highest BEC value into four categories: (1) loops, all with lag-lengths 42 or 43, (2) edges from global radiation intensity into all of the gases, with lag-lengths between 5 and 8, (3) edges from one gas into another, with lag-length either one or two, and (4) edge from NO into GRI. We notice that expert knowledge from a scientist working with at-

Table 5 Marginal distributions for the lag lengths of the directed edges of the Granger causality graph for the air pollution data. Only the most probable lag lengths are shown, along with their probabilities, such that the total probability of the shown values exceeds 0.99 for each edge

	CO	NO	NO_2	O_3	GRI
CO	43: 0.99 44: 0.01	3: 0.72 1: 0.24 6: 0.03	0: 1.00	1: 0.53 2: 0.47	0: 1.00
NO	1: 0.97 6: 0.03	43: 0.97 44: 0.03	1: 1.00	0: 1.00	2: 0.95 3: 0.05
NO_2	2: 0.52 1: 0.48	2: 0.50 1: 0.40 0: 0.10	42: 0.98 43: 0.02	2: 1.00	0: 1.00
O_3	1: 0.98 0: 0.02	0: 0.98 1: 0.02	1: 1.00	43: 1.00	0: 1.00
GRI	7: 0.96 3: 0.04	5: 0.88 7: 0.08 4: 0.04	5: 0.81 6: 0.18	8: 0.96 7: 0.04	42: 0.98 37: 0.02

mospheric gases would be required for a proper interpretation of the results. However, some simple conclusions can be made, but we wish to emphasize that the analysis presented here is based on common sense reasoning. Lag-length 42 corresponds to a period of seven days, reflecting a weekly cycle for the whole process. This can be explained by different traffic conditions on different days of the week, especially the asymmetry between the weekends and weekdays. Notice also the local optima in the distributions of the lag-lengths of the loops, which are clearly visible in Fig. 3 with a period of one day. The weekly cycle gets clear support when the relative heights of two consecutive local optima are compared. It is seen that the difference between the local optima corresponding to the seventh and the sixth day is clearly larger than differences between the previous local optima. This is especially evident in the plots corresponding to CO and NO. The directed edges from GRI into the gases with lag-length 5–8 indicate that the weather conditions are Granger-causal to the level of pollutants with one day's lag. The interactions between different pollutants are more short-term, indicated by the shorter lag-lengths of the corresponding directed edges. An interesting anomaly is the edge from NO into global radiation level with lag-length less than half a day, being the only directed edge entering GRI. It is possible that instead of representing a true interaction between NO and GRI (which would appear unlikely), this edge may represent some latent underlying process, in which the coming weather conditions are anticipated, and actions are made accordingly.

In order to investigate the effect of considering the lag-length of each edge separately, we ran the search algorithm with the constraint that each directed edge must have a specified lag-length. The constrained version of the algorithm was run with all lag-length values from 1 to 60. Again, BEC values showed a clear cyclic behavior, and local optimas were obtained in daily intervals (exact values are not shown). The directed part of the constrained graph with the highest BEC-score is shown in Fig. 1b, and it corresponds to lag-length 43, highlighting the weekly characteristic of the process. For comparison, the directed part of the graph corresponding to the smallest locally optimal lag-length value (equal to 7) is shown in Fig. 1c. The undirected parts of both of these graphs were equal to the undirected

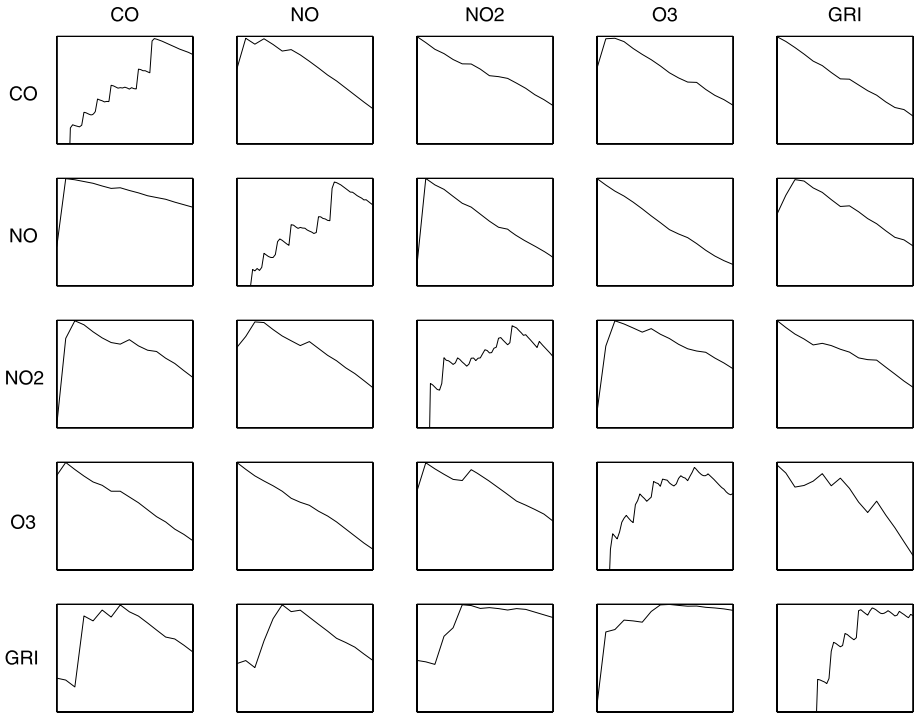


Fig. 3 Conditional log posterior distributions (8) of the lag lengths of the directed edges of the Granger causality graph for the air pollution data, given the graph with the highest BEC value. The limits of the x-axes of the images on the diagonal, representing the loops, correspond to lag lengths of 0 and 60 observations, corresponding to a time span of ten days. The x-axes of all other images are limited by lag lengths of 0 and 15 observations, corresponding to a period of 2.5 days. The y-axes are shown on a logarithmic scale and the limits are different for different images, chosen to best illustrate the shape of the curve

graph of the globally optimal graph, shown in Fig. 2a. In contrast, the directed part of the graph changed dramatically depending on whether unequal lag-lengths were allowed, as the optimal constrained graph in Fig. 1b has only one edge, in addition to the loops. As with the simulated examples in Sect. 4.1, many directed edges with lower lags were omitted. An interesting feature in the graph in Fig. 1c, corresponding to lag-length 7, is that it has more directed edges than the optimal directed graph in Fig. 1a. Especially, there are now three directed edges entering GRI, as compared to the optimal graph where the edge from NO is the only one to enter GRI. We conjecture that the additional edges could be explained as a compensation for the underestimated memories of the processes: an edge from a process into another process, from which there is an edge back into the original process, may act as a longer memory for the first process. However, we did not observe such behavior in the simulated examples.

The estimation of the undirected part of the graph seems to be quite independent of the estimation of the directed edges, since the optimal graph and the other two graphs presented in Fig. 1 all have the same undirected graph, shown in Fig. 2a. It is intriguing to compare the undirected parts of the graph obtained from the current method, and the one in Fig. 2b obtained from Corander and Villani (2006). The main difference between the two methods is that here we are not forcing decomposability to the undirected graph. Notice that the graph

estimated by the presented method in Fig. 2a includes a chordless four-cycle (CO, NO, O_3, NO_2) and is therefore non-decomposable, while the graph in Fig. 2b is decomposable. Figure 2c shows the undirected edges of a graph which was obtained by Corander and Villani, when they used pairwise testing to detect the edges i.e. for each edge the complete graph was compared to a graph where the edge was missing, and, if the graph with the missing edge got higher Bayesian score values, the edge was excluded. Such a conservative strategy may lead to an overly simplified model, as illustrated by Corander and Villani, and can be seen by comparing Figs. 2c to 2a and 2b. Surprisingly, edge NO-GRI is not present in Fig. 2b although it is present in both Figs. 2a and 2c. Also Dahlhaus and Eichler (2003) found NO and GRI to be contemporaneously partially correlated. This missing edge in Fig. 2b could be due to the fact that the corresponding analysis was constrained to include decomposable graphs only.

To investigate the predictive ability under different constraints with realistic data, we compared VAR(G, L) and VAR($\widetilde{G}$) models in two different situations (VAR(G, p^*) was omitted from this comparison due to its clearly lower predictive ability observed in the simulations). First, we predicted 100 last observations from the series using the model and ml-estimates learned from all the preceding data (4286 observations). Because the scale of the different components in the series vary radically, we calculated the MSE separately for each of the components. We summarize the results by noting that four out of five components had smaller MSE values when the unconstrained model VAR($\widetilde{G}$) was used. As the second scenario, we used 300 observations from the beginning of the series to learn the model and the ml-estimates and used these to predict the following 10 observations. In this case, with only a limited amount of data available, the results became reversed, such that four out of five component MSEs were smaller with the VAR(G, L) model than with the unconstrained VAR($\widetilde{G}$) model.

The presented analyses were run on a standard desktop PC with a 2.2 GHz processor. With our current MATLAB implementation, one iteration of the search took about three minutes. Thus, one run of the search algorithm (16 iterations) for the air pollution data was completed approximately within an hour. As discussed in Sect. 3.2, the running time depends both on the dimension of the process, as well as the complexity of the underlying interactions. For comparison, we ran the search algorithm with a graph, which was simulated using the parameters of the sparse graph type (rows 4–6 in Table 1), except that the dimension was specified to be 20. Then, the time required for one iteration was slightly less than with the air pollution data with five dimensions (exact results not shown). These examples illustrate that the applicability depends in practice largely on the complexity of the problem. If the underlying graph is relatively sparse, and the lag-lengths are relatively small, say up to 5, then we expect the presented algorithm to work well with graphs with at least 20–30 nodes. On the other hand, if there are numerous edges with long lags, then even a graph with 10 nodes may provide a serious challenge to the algorithm from the perspective of time complexity. Nevertheless, even such situations could be more easily handled by using parallel computing to perform the search repeatedly on separate CPUs. Also ideas from adaptive MCMC methods (e.g. Haario et al. 2001) could be used to fine tune the algorithm further. In adaptive MCMC the search process is able to adapt to the most promising areas in the model space. In fact, a naive version of such an approach was implicitly used also in our analysis, when we specified a different upper bound for the lag-lengths of loops and other directed edges, once the preliminary tests had shown that such a strategy would sufficiently cover all the promising areas of the model space.

To further illustrate the benefit of using greedy iterations in between the stochastic ones, Fig. 4 shows a trace plot of BEC values of the models visited during a search, which was

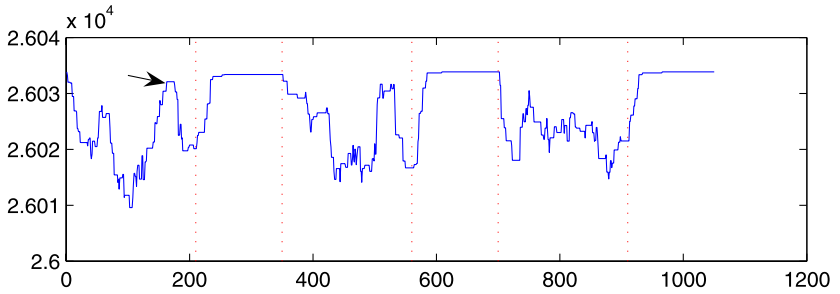


Fig. 4 Trace plot of BEC values of models visited during a search which was started from the optimal model for the air pollution data and carried out by repeating three times a sequence of six stochastic and four greedy iterations. The periods of stochastic and greedy iterations are separated by *dotted vertical lines*. The highest score found during the stochastic iterations is marked with an *arrow* and it is 5.7 times less probable than the overall optimal model

started from the found optimal model, and accomplished by repeating a sequence of 6 stochastic and 4 greedy iterations three times. As can be observed by investigation of Fig. 4, the use of stochastic iterations immediately led to models with lesser predictive ability, while the greedy iterations very rapidly found their way to the optimal model. In fact, the model with the highest BEC-score found during the stochastic iterations was 5.7 times less likely than the global optimal model, and, in total 10 models with higher BEC-scores were visited during the greedy iterations. Thus, the models visited during the stochastic iterations had in this case only a small effect to the calculated posterior probabilities, even if the stochastic iterations are important for a thorough exploration of the model space, as discussed in Sect. 3.2.

Finally, we briefly summarize the differences between the presented analysis with the original data and the analysis with the trend-corrected residual data, for which detailed results are skipped. We remark that when Dahlhaus (2000) analyzed the same data using either raw or residual data, both the analyses yielded the same partial correlation graph. Also in our analyses the general picture with the residual data was quite similar to the original data, namely, most of the edges were present in the graph and the loops had noticeably larger lag-lengths than other directed edges. The main difference to the presented analysis was that the lags were clearly shorter for the trend-corrected data. The lags for the loops were around 50, corresponding to a period of about one day (recall that the analyzed residual data contained 48 observations per day, as opposed to 6 per day in the presented raw data analysis), while other edges had lags less than 8. Thus, the weekly periodicity observed with the raw data was removed by focusing the analysis to the residuals. The residual analysis also included a few more edges in the estimated optimal graph than the raw data analysis. However, the directed edges which most obviously should be excluded from the graph, namely, those from the air pollutants to the weather conditions (GRI), were still excluded, apart from the edge from NO to GRI which was included in both the analyses. When we analyzed the residual data using the constraint of equal lag-lengths, similar behavior as with the original data was observed: the dynamic part of the graph changed considerably as compared to the unconstrained analysis, while the undirected part of the graph remained unaffected.

5 Discussion

The recent interest shown by scientists working with functional neuroimaging and critical care monitoring systems shows that fruitful insights to dynamic dependence structures are achievable by appropriately sparse vector autoregressive models, which exploit graphical modelling strategies. Our analyses of both the synthetic and the air pollution data demonstrated the importance of not restricting the core characteristics of the considered model class due to mathematical and computational convenience, as this may lead to misleading structural representations. This is particularly relevant as the graphs themselves offer an intuitive interface towards Granger causality.

Our experiments with simulated and empirical data demonstrate that gains in predictive accuracy may be achieved by considering structural learning of vector autoregressive processes, as compared to learning with unstructured models. The benefits become accentuated especially when the amount of data to estimate the model and the parameters is limited. The unstructured models represent a very flexible class of probabilistic machinery that can be used for exploring dynamic systems. However, as illustrated by the simulations, the unstructured models also easily capture local random patterns in the data and convert these into spurious causality conclusions. Therefore, it would be of interest to perform in future experiments on a larger scale, to compare the predictive performance of structured Granger causality graphs with that of standard dynamic Bayesian network models. In particular, such studies could shed more light on the typical level of sparsity achievable in a Granger causality graph representation of real multivariate dynamic systems.

Our investigations also show that the introduced learning algorithm using a combination of stochastic and greedy iterations to optimize the derived Bayesian model selection criterion performs competently when confronted with the extreme size of the space of all VAR(G, L) models in a realistic scenario. However, the vast dimensionality of many fMRI data sets would still require further elaboration of the learning algorithm introduced here, as well as suggest the possibility of utilizing sparser parametrizations, such as the graphical factor analysis models introduced in Giudici and Stanghellini (2002). Similar ideas were independently considered in the neuroimaging work by Valdés-Sosa et al. (2005). Another potential direction of future development would be to merge the graphical modelling strategy with the type of parametric characterization used in co-integration models for time series (Johansen 1995).

Appendix

A.1 Derivation of BEC

Here the expected utility

$$\overline{u}(M_j|\mathbf{x}) = \int_{\Theta_j} u(M_j, \theta_j)\pi(\theta_j|\mathbf{x})d\theta_j \tag{11}$$

is derived for a Granger causality graph G and lag matrix L of a VAR(G, L) process. Let $\mathcal{C}$ and $\mathcal{S}$ denote the sets of mp-components and separators in the undirected part of G. Consider

a VAR(G, L) model, with upper bound p for the lag-lengths of the processes:

$$x_t = B_1 x_{t-1} + \cdots + B_p x_{t-p} + \epsilon_t, \quad t = 1, \ldots, n, \tag{12}$$

where x_i, are $(k \times 1)$ vectors and B_i $(k \times k)$ coefficient matrices, $i = 1, \ldots, p$, and x_i, $i = -p+1, \ldots, 0$ are assumed to be available. Here k represents the dimension of the process. The vectors ϵ_i are assumed to be i.i.d. $N(0, \Sigma)$. The system of (12) can be rewritten using matrix notation as

$$Y = XB + U, \tag{13}$$

where

$$\begin{aligned}
Y(n \times k) &= [x_1, x_2, \ldots, x_n]^T, \\
X(n \times kp) &= \begin{bmatrix} x_0^T & x_{-1}^T & \cdots & x_{-p+1}^T \\ x_1^T & x_0^T & \cdots & x_{-p+2}^T \\ \vdots & \vdots & \vdots & \vdots \\ x_{n-1}^T & x_{n-2}^T & \cdots & x_{n-p}^T \end{bmatrix}, \\
B(kp \times k) &= [B_1, B_2, \ldots, B_p]^T, \quad \text{and} \\
U(n \times k) &= [\epsilon_1, \ldots, \epsilon_n]^T.
\end{aligned}$$

Given the facts that ϵ_i are i.i.d. $N(0, \Sigma)$, and that the distribution of ϵ_i factorizes according to the decomposition of G into mp-components, and using the known expression for the entropy of a multivariate Gaussian distribution (e.g. Whittaker 1990), the utility $u(M_j, \theta_j)$ can now be written as:

$$\begin{aligned}
u(M_j, \theta_j) &= u(M_j, (\Sigma, B)) \\
&= \int \Pr(U|B, \Sigma) \log \Pr(U|B, \Sigma) dU \\
&= n \int \Pr(\epsilon|B, \Sigma) \log \Pr(\epsilon|B, \Sigma) d\epsilon \\
&= n \sum_{c \in \mathcal{C}} \int \Pr(\epsilon_c|B, \Sigma_c) \log \Pr(\epsilon_c|B, \Sigma_c) d\epsilon_c \\
&\quad - n \sum_{s \in \mathcal{S}} \int \Pr(\epsilon_s|B, \Sigma_s) \log \Pr(\epsilon_s|B, \Sigma_s) d\epsilon_s \\
&= n \sum_{s \in \mathcal{S}} \left(\frac{|s|}{2} + \frac{|s|}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_s| \right) \\
&\quad - n \sum_{c \in \mathcal{C}} \left(\frac{|c|}{2} + \frac{|c|}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_c| \right),
\end{aligned} \tag{14}$$

where ϵ is a random variable from $N(0, \Sigma)$, and ϵ_c and ϵ_s are subvectors of ϵ containing the elements in prime component c, or separator s, respectively. Here $|\cdot|$ is used both for the determinant of a matrix and the cardinality of a set. It follows from the form of (14)

that the computation of (11) can be divided according to the prime components and the corresponding separators of the graph such that:

$$\overline{u}(M_j|\mathbf{x}) = \sum_{c\in\mathcal{C}} \overline{u}(M_{j,c}|\mathbf{x}) - \sum_{s\in\mathcal{S}} \overline{u}(M_{j,s}|\mathbf{x}), \tag{15}$$

where $M_{j,c}$ is a submodel corresponding to the elements in c. Since (14) does not depend on B, $\overline{u}(M_{j,c}|\mathbf{x})$ can be obtained by integrating with respect to the marginal posterior distribution of Σ_c:

$$\overline{u}(M_{j,c}|\mathbf{x}) = -n \int \left(\frac{|c|}{2} + \frac{|c|}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_c| \right) \Pr(\Sigma_c|Y, X) d\Sigma_c. \tag{16}$$

To derive the posterior $\Sigma_c|Y, X$, assume first that prime component c of the undirected part of G is a clique, and that there are no restrictions for the entries of B_c (matrix consisting of columns of B corresponding to the elements of c). In this case, a derivation from the reference priors

$$\Pr(B_c) = const, \quad \text{and}$$
$$\Pr(\Sigma_c) \propto |\Sigma_c|^{-(k+1)/2},$$

to the posterior

$$\Sigma_c|X, Y_c \sim W_{|c|}^{-1}(S_c^{-1}, n - kp), \tag{17}$$

can be found in Zellner (1971). Thus, the posterior (17) is an inverse Wishart distribution with expected value proportional to S_c, and $n - kp$ degrees of freedom, where

$$S_c = (Y_c - X\widehat{B}_c)^T (Y_c - X\widehat{B}_c),$$

and

$$\widehat{B}_c = (X^T X)^{-1} X^T Y_c,$$

a matrix of least squares values. When there are missing edges causing zero restrictions on the elements of B_c, the resulting posterior distribution is not a standard distribution. Furthermore, if c is not a clique, but has some missing edges resulting in zero restrictions on elements of Σ^{-1}, the corresponding distribution is known as Hyper Inverse Wishart, which is discussed in the context of Bayesian inference for Gaussian non-decomposable graphical models in Roverato (2002). Taking these restrictions into account in an exact manner would make the analytical evaluation of (16) impossible. However, to be able to utilize an analytical evaluation of (16), we approximate the posterior distribution by

$$\Sigma_c|X, Y_c \sim W_{|c|}^{-1}\left(S_c^{-1}, n - \frac{\beta_c}{|c|} + \frac{\gamma_c}{|c|}\right), \tag{18}$$

where β_c is the number of unrestricted parameters in B_c, and γ_c represents the number of missing edges between the nodes of mp-component c. Thus, the approximation has the same expected value as the corresponding unrestricted case, but the increased (or decreased) uncertainty caused by increased (decreased) number of parameters is taken into account in the decreased (increased) degrees of freedom. The main justification for this approximation is that it preserves the asymptotically consistent behavior of the criterion, as will be shown

later. Approximation (18) is exact if c is a clique and either there is a directed edge from every node to all other nodes within c, or there are no directed edges at all entering the nodes in c.

Using Lemma 5.1 in Corander (2003), the integration in (16) can now be performed and the expected utility related to the prime component c equals:

$$\overline{u}(M_{j,c}|\mathbf{x}) = -\frac{n|c|}{2}(1+\log\pi+\log n) - \frac{n}{2}\log|\widehat{\Sigma}_c| + \frac{n}{2}\sum_{i=0}^{|c|-1}\psi\left(\frac{\nu_c - i}{2}\right), \tag{19}$$

where $\widehat{\Sigma}_c$ is the maximum likelihood estimate of the covariance matrix, ψ is the digamma function and

$$\nu_c = n - \frac{\beta_c}{|c|} + \frac{\gamma_c}{|c|}$$

is the degrees of freedom in the posterior distribution of Σ_c. The expected utility $\overline{u}(M_{j,s}|\mathbf{x})$ for the separators $s \in \mathcal{S}$ in (15) can be calculated analogously.

The maximized log-likelihood of a Gaussian model (see Mardia et al. 1979, Chap. 4.2) is given by

$$\log p(x|\widehat{\theta}) = -\frac{n}{2}\log|2\pi\,\widehat{\Sigma}| - \frac{nk}{2}.$$

Since $|\widehat{\Sigma}|$ factorizes according to the mp-components, the insertion of terms of type (19) to (15) leads to the following form of the expected utility for the model:

$$\overline{u}(M_j|\mathbf{x}) = \log p_j(x|\widehat{\theta}_j) + f(M_j),$$

where

$$f(M_j) = -\frac{nk}{2}\log\left(\frac{n}{2}\right) + \frac{n}{2}\sum_{c\in\mathcal{C}}\sum_{i=0}^{|c|-1}\psi\left(\frac{\nu_c - i}{2}\right) - \frac{n}{2}\sum_{s\in\mathcal{S}}\sum_{i=0}^{|s|-1}\psi\left(\frac{\nu_s - i}{2}\right),$$

can be evaluated analytically.

Finally, we prove a lemma concerning the asymptotic properties of the difference $f(M_j) - f(M_l)$:

Lemma 2 *$f(M_j) - f(M_l) \overset{n\to\infty}{\to} \frac{d_l - d_j}{2}$, where d_j and d_l denote the number of free parameters in the two models.*

Proof Using an asymptotic approximation (Abramovitz and Stegun 1965) for the digamma function ψ:

$$\psi(x) \approx \ln x - \frac{1}{2x},$$

the difference $f(M_j) - f(M_l)$ can be written as:

$$\begin{aligned}
&\frac{n}{2}\ln\left(\frac{\prod_{c_j\in\mathcal{C}_j}\prod_{i=0}^{|c_j|-1}(n-\frac{\beta_{c_j}}{|c_j|}+\frac{\gamma_{c_j}}{|c_j|}-i)}{\prod_{s_j\in\mathcal{S}_j}\prod_{i=0}^{|s_j|-1}(n-\frac{\beta_{s_j}}{|s_j|}-i)}\right)\\
&\quad-\frac{n}{2}\ln\left(\frac{\prod_{c_l\in\mathcal{C}_l}\prod_{i=0}^{|c_l|-1}(n-\frac{\beta_{c_l}}{|c_l|}+\frac{\gamma_{c_l}}{|c_l|}-i)}{\prod_{s_l\in\mathcal{S}_l}\prod_{i=0}^{|s_l|-1}(n-\frac{\beta_{s_l}}{|s_l|}-i)}\right)+O(n^{-1})\\
&=\frac{n}{2}\ln\left(\frac{n^X-An^{X-1}+O(n^{X-2})}{n^X-Bn^{X-1}+O(n^{X-2})}\right)+O(n^{-1}),
\end{aligned}\tag{20}$$

where

$$\begin{aligned}
X&=\sum_{c_j\in\mathcal{C}_j}|c_j|+\sum_{s_l\in\mathcal{S}_l}|s_l|=\sum_{c_l\in\mathcal{C}_l}|c_l|+\sum_{s_j\in\mathcal{S}_j}|s_j|,\\
A&=\sum_{c_j\in\mathcal{C}_j}\left(\beta_{c_j}-\gamma_{c_j}+\frac{(|c_j|-1)|c_j|}{2}\right)+\sum_{s_l\in\mathcal{S}_l}\left(\beta_{s_l}+\frac{(|s_l|-1)|s_l|}{2}\right),
\end{aligned}\tag{21}$$

and B is obtained from A by switching j and l. The second equality in (21) follows from the fact that $\sum_{c_l\in\mathcal{C}_l}|c_l|-\sum_{s_l\in\mathcal{S}_l}|s_l|=\sum_{c_j\in\mathcal{C}_j}|c_j|-\sum_{s_j\in\mathcal{S}_j}|s_j|=k$. Continuing from (20), $f(M_j)-f(M_l)$ equals:

$$\begin{aligned}
&\frac{n}{2}\ln\left(1+\frac{(B-A)n^{X-1}+O(n^{X-2})}{n^X-Bn^{X-1}+O(n^{X-2})}\right)+O(n^{-1})\\
&\approx\frac{n}{2}\frac{(B-A)n^{X-1}+O(n^{X-2})}{n^X-Bn^{X-1}+O(n^{X-2})}+O(n^{-1})\quad\text{(for large } n)\\
&=\frac{(B-A)+O(n^{-1})}{2+O(n^{-1})}+O(n^{-1})\overset{n\to\infty}{\to}\frac{B-A}{2}.
\end{aligned}\tag{22}$$

Because the number of free parameters in the coefficient matrices of, say model M_j, is given by

$$\sum_{c_j\in\mathcal{C}_j}\beta_{c_j}-\sum_{s_j\in\mathcal{S}j}\beta_{s_j},$$

and the number of free parameters in the precision matrix is given by

$$\sum_{c_j\in\mathcal{C}_j}\left(\frac{(|c_j|-1)|c_j|}{2}-\gamma_{c_j}\right)-\sum_{s_j\in\mathcal{S}j}\frac{(|s_j|-1)|s_j|}{2}+k,$$

it can be seen that $B-A$ in (22) represents the difference in the number of the free parameters of the two models, as required. □

It follows that BEC criterion is asymptotically equivalent to the criterion of Hannan and Quinn (1979), as stated in Sect. 3.1.

A.2 Maximum likelihood estimation of the model parameters

Here we describe an algorithm for the calculation of the maximized log-likelihood $\log p_j(\mathbf{x}|\widehat{\theta}_j)$. The maximum likelihood (ml) estimates of the coefficient matrices B_i, $i = 1, \ldots, p$, and the residual covariance matrix Σ under restrictions imposed by a Granger causality graph G and lag matrix L, are obtained by repeatedly performing the following two steps until convergence:

1. Conditional on the current estimate of Σ, compute the ml-estimates of B_i, $i = 1, \ldots, p$.
2. Conditional on the current estimates of B_i, $i = 1, \ldots, p$, compute the ml-estimate of Σ.

The above algorithm belongs to the class of iterative partial maximization algorithms, discussed in Lauritzen (1996, Appendix A.4). A similar algorithm was recently used by Drton and Eichler (2006) for the ml-estimation in Gaussian chain graph models. The algorithm is guaranteed to converge to a point in which the gradient of the log-likelihood equals zero. Assuming that this point represents a global maximum, the algorithm thus converges. However, the existence of a unique optimum can not be proven. In the related context discussed by Drton and Eichler, the likelihood may indeed be multimodal, which makes it likely that this could be the case in the current setting also, although we have not investigated this possibility further. However, the multimodality mentioned by Drton and Eichler is limited to cases with a very limited number of observations, and seems to have no consequence in practice. For further discussion on the topic, see Drton and Eichler (2006) and the references therein.

The updating operation for step 1 is given in Lütkepohl (1993, Formula 5.2.17). To perform the second step of the maximum likelihood estimation, the error terms ϵ_t in (1) are first calculated, conditional on the current estimates of B_i. Because these error terms are independent realizations from a multinormal distribution with Σ as the covariance matrix, the theory from Lauritzen (1996, Chap. 5) is directly applicable. The undirected part of G is divided into maximal prime components. Algorithm for this is given e.g. in Leimer (1993). Proposition 5.9 from Lauritzen for ml-estimation of decomposable Gaussian models generalizes straightforwardly to ml-estimation for models decomposed into their mp-components. The ml-estimate of the precision matrix $\widehat{K}$ (inverse of $\widehat{\Sigma}$) can be obtained from the ml-estimates of the precision matrices corresponding to the mp-components of the graph according to:

$$\widehat{K} = \sum_{c \in \mathcal{C}} [\widehat{K}_c]^c_{k \times k} - \sum_{s \in \mathcal{S}} \alpha(s) [\widehat{K}_s]^s_{k \times k} \tag{23}$$

where $\mathcal{C}$ is the set of all mp-components and $\mathcal{S}$ is the corresponding set of separators with multiplicities α in the sequence of the used decompositions. The operator $[\cdot]^c_{k \times k}$ denotes a $k \times k$ matrix, whose submatrix corresponding to the elements of c is given inside the parentheses and other elements are zeros. For details, see Lauritzen (1996, Sect. 5.3.2). If mp-component c is a clique, $\widehat{\Sigma}_c$ is given by a submatrix of the sample covariance matrix of the error terms, corresponding to the elements in c (Lauritzen 1996, Theorem 5.1). Otherwise iterative methods must be used. For computing $\widehat{K}_c$ of non-complete mp-components we used the basic form of iterative proportional fitting, which is described in Speed and Kiiveri (1986), along with some alternatives.

References

Abramovitz, M., & Stegun, I. A. (Eds.) (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, *21*, 243–247.
Bach, F. R., & Jordan, M. I. (2004a). Beyond independent components: Trees and clusters. *Journal of Machine Learning Research*, *4*, 1205–1233.
Bach, F. R., & Jordan, M. I. (2004b). Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, *52*, 2189–2199.
Bernardo, J. M. (1999). Nested hypothesis testing: the Bayesian reference criterion. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 6, pp. 101–130). London: Oxford University Press. With discussion.
Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester: Wiley.
Brillinger, D. R. (1996). Remarks concerning graphical models for time series and point processes. *Revista de Econometria*, *16*, 1–23.
Brüggemann, R., Krolzig, H.-M., & Lütkepohl, H. (2002). Comparison of model reduction methods for VAR processes. EUI Working Paper, ECO, 2002/19. http://hdl.handle.net/1814/791.
Carvalho, C., & West, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis*, *2*, 69–98.
Corander, J. (2003). Bayesian graphical model determination using decision theory. *Journal of Multivariate Analysis*, *85*, 253–266.
Corander, J., & Marttinen, P. (2006). Bayesian model learning based on predictive entropy. *Journal of Logic, Language and Information*, *15*, 5–20.
Corander, J., & Villani, M. (2006). A Bayesian approach to modelling graphical vector autoregressions. *Journal of Time Series Analysis*, *27*, 141–156.
Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (2001). *Introduction to algorithms* (2nd edn.). Cambridge: MIT Press.
Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, *51*, 157–172.
Dahlhaus, R., & Eichler, M. (2003). Causality and graphical models in time series analysis. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 115–137). London: Oxford University Press.
Dash, D. (2005). Restructuring dynamic causal systems in equilibrium. In: R. Cowell & Z. Ghahramani (Eds.), *Proceedings of the tenth international workshop on artificial intelligence and statistics (AIStats)*. Society for artificial intelligence and statistics. Available electronically at http://www.gatsby.ucl.ac.uk/aistats/.
Drton, M., & Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics*, *33*, 247–257.
Eichler, M. (2001). *Graphical modelling of multivariate time series*. Technical report, Universität Heidelberg. arXiv:math.ST/0610654.
Eichler, M. (2006a). Fitting graphical interaction models to multivariate time series. In *Proceedings of the 22nd conference of uncertainty in artificial intelligence*. Arlington: AUAI Press.
Eichler, M. (2006b). Graphical modelling of dynamic relationships in multivariate time series. In M. Winterhalder, B. Schelter, & J. Timmer (Eds.), *Handbook of time series analysis* (pp. 335–372). New York: Wiley.
Eichler, M. (2007). Granger-causality and path diagrams for multivariate time series. *Journal of Econometrics*, *137*, 334–353.
Eichler, M. (2008). Causal inference from multivariate time series: What can be learned from Granger causality. In C., Glymour, W. Wang & D. Westerstahl (Eds.), *Proceedings from the 13th international congress of logic, methodology and philosophy of science*. King's College Publications, London.
Eichler, M., Dahlhaus, R., & Sandkühler, J. (2003). Partial correlation analysis for the identification of synaptic connections. *Biological Cybernetics*, *89*, 289–302.
Florens, J. P., & Mouchart, M. (1985). A linear theory for noncausality. *Econometrica*, *53*, 157–175.
Fried, R., & Didelez, V. (2003). Decomposability and selection of graphical models for multivariate time series. *Biometrika*, *90*, 251–267.
Fried, R., & Didelez, V. (2005). Latent variable analysis and partial correlation graphs for multivariate time series. *Statistics & Probability Letters*, *73*, 287–296.
Friedman, N., Murphy, K., & Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In G. F. Cooper & S. Moral (Eds.), *Proceedings of the 14th annual conference on uncertainty in artificial intelligence (UAI-98)*. San Mateo: Morgan Kaufmann.
Gather, U., Imhoff, M., & Fried, R. (2002). Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, *21*, 2685–2701.
Giudici, P., & Stanghellini, E. (2002). Bayesian inference for graphical factor analysis models. *Psychometrika*, *66*, 577–592.
Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*, 24–36.

Granger, C. W. J. (2001). *Essays in econometrics: collected papers of Clive W.J. Granger*. Cambridge: Cambridge University Press. Ghysels, E., Swanson, N.R. & Watson, M.W. (Eds.).
Gredenhoff, M., & Karlsson, S. (1999). Lag-length selection in VAR-models using equal and unequal lag-length procedures. *Computational Statistics*, *14*, 171–187.
Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, *7*, 223–242.
Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, *B 41*, 190–195.
Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks—the combination of knowledge and statistical data. *Machine Learning*, *20*, 197–243.
Imhoff, M., & Kuhls, S. (2006). Alarm algorithms in critical care monitoring. *Anesthesia and Analgesia*, *102*, 1525–1537.
Iwasaki, Y., & Simon, H. A. (1994). Causality and model abstraction. *Artificial Intelligence*, *67*, 143–194.
Janzura, M., & Nielsen, J. (2006). A simulated annealing-based method for learning Bayesian networks from statistical data. *International Journal of Intelligent Systems*, *21*, 335–348.
Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. London: Oxford University Press.
Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.
Koivisto, M., & Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, *5*, 549–573.
Lauritzen, S. L. (1996). *Graphical models*. London: Oxford University Press.
Leimer, H.-G. (1993). Optimal decomposition by clique separators. *Discrete Mathematics*, *113*, 99–123.
Lütkepohl, H. (1993). *Introduction to multiple time series analysis*. Berlin: Springer.
Lynggaard, H., & Walther, K. H. (1993). *Dynamic modelling with mixed graphical association models*. Master's thesis, Aalborg University.
Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. San Diego: Academic Press.
Moneta, A., & Spirtes, P. (2005). Graph-based search procedure for vector autoregressive models. LEM Working Paper 2005/14, Sant'Anna School of Advanced Studies, Pisa.
Oxley, L., Reale, M., & Tunnicliffe, W. (2004). Finding directed acyclic graphs for vector autoregressions. In J. Antoch (Ed.), *Proceedings in computational statistics 2004* (pp. 1621–1628). Heidelberg: Physica.
Ozcicek, O., & McMillin, W. D. (1999). Lag length selection in vector autoregressive models: symmetric and asymmetric lags. *Applied Economics*, *31*, 517–524.
Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.
Reale, M., & Tunnicliffe Wilson, G. (2001). Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications*, *10*, 49–65.
Reale, M., & Tunnicliffe Wilson, G. (2002). The sampling properties of conditional independence graphs for structural vector autoregressions. *Biometrika*, *8*, 457–461.
Robert, C. P., & Casella, G. (2005). *Monte Carlo statistical methods* (2nd ed.). New York: Springer.
Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, *29*, 391–411.
Salvador, R., Suckling, J., Schwarzbauer, C., & Bullmore, E. (2005). Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions of the Royal Society B Biological Sciences*, *360*, 937–946.
Schelter, B., Winterhalder, M., Hellwig, B., Guschlbauer, B., Lucking, C. H., & Timmer, J. (2006). Direct or indirect? Graphical models for neural oscillators. *Journal of Physiology (Paris)*, *99*, 37–46.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
Seinfeld, J. H. (1986). *Atmospheric chemistry and physics of air pollution*. New York: Wiley.
Sisson, S. A. (2005). Transdimensional Markov chains: a decade of progress and future perspectives. *Journal of the American Statistical Association*, *100*, 1077–1089.
Speed, T. P., & Kiiveri, H. T. (1986). Gaussian distributions over finite graphs. *Annals of Statistics*, *14*, 138–150.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge: MIT Press.
Stanghellini, E., & Whittaker, J. (1999). Analysis of multivariate time series via a hidden graphical model. In D. Heckerman & J. Whittaker (Eds.), *Proceedings of the seventh international workshop on artificial intelligence and statistics*. San Mateo: Morgan Kaufmann.
Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., & Canalez-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *360*, 969–981.
Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. New York: Wiley.
Winker, P., & Maringer, D. (2004). Optimal lag structure selection in VEC-models. Computing in Economics and Finance 2004 155, Society for Computational Economics.
Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley.