Learning the Structure of a Nonstationary Vector Autoregression

Daniel Malinsky
Department of Computer Science
Johns Hopkins University
Baltimore, MD USA
malinsky@jhu.edu

Peter Spirtes
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA USA
ps7z@andrew.cmu.edu

Abstract

We adapt graphical causal structure learning methods to apply to nonstationary time series data, specifically to processes that exhibit stochastic trends. We modify the likelihood component of the BIC score used by score-based search algorithms, such that it remains a consistent selection criterion for integrated or cointegrated processes. We use this modified score in conjunction with the SVAR-GFCI algorithm [15], which allows us to recover qualitative structural information about the underlying data-generating process even in the presence of latent (unmeasured) factors. We demonstrate our approach on both simulated and real macroeconomic data.

1 Introduction

Standard methods for causal structure learning from time series data, including techniques based on graphical models, typically rely on a stationarity assumption. A multivariate stochastic process is nonstationary when the joint distribution (or conditional distribution, given initial values) of the variables changes over time. There are many varieties of nonstationarity, but the present focus will be on "stochastic trends," which are common features of (e.g.) macroeconomic time series, such as interest and inflation rates. Stochastic trends occur when the usual autoregressive stability condition fails, and so some variables exhibit so-called "random walk" behavior [14]. Specifically, we consider the case where the characteristic polynomial has some unit roots, but the process becomes stable upon differencing (so, we exclude explosive processes). This includes cointegrated processes, wherein linear combinations of multiple time series are stationary.

We begin with some brief background on structural vector autoregressions, causal graphical models, and algorithms which learn causal graphs from time series data. Then, we use a toy example to explain why simply taking first differences of nonstationary variables is not sufficient for accurate structure learning, and instead we introduce an approach which makes use of a reparameterization of the data-generating process in error-correction form. Finally, we present some simulation results and an application to macroeconomic time series data from Denmark to support our claim that structure learning from nonstationary data can be feasibly and fruitfully tackled by graphical methods.

2 Background

In the following, we use boldface letters to denote vectors of random variables, e.g., $\mathbf{X}_t$ is a $k \times 1$ vector of time series variables $(X_{1,t}, ..., X_{k,t})'$. To accomodate the possibility of unmeasured variables, we sometimes write $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$ where $\mathbf{L}_t$ is a $l \times 1$ vector of latent components.

2.1 Structural VAR models and their graphical representations

Causal models are often presented as systems of nonparametric structural equations, which have corresponding graphical representations [23, 18, 19]. In our dynamical setting, we assume the underlying stochastic process $\{\tilde{\mathbf{X}}_t\}_{t \in \mathbb{N}}$ is generated by some unknown linear structural vectorautoregression (SVAR), including arbitrary latent components. A $(k+l)$-dimensional order-p SVAR process may be written:

$$\tilde{\mathbf{\Gamma}}_0 \tilde{\mathbf{X}}_t = \tilde{\mathbf{\Gamma}}_1 \tilde{\mathbf{X}}_{t-1} + ... + \tilde{\mathbf{\Gamma}}_p \tilde{\mathbf{X}}_{t-p} + \tilde{\boldsymbol{\varepsilon}}_t \qquad (1)$$

$\forall t \in \mathbb{N}$ where $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$, $\tilde{\boldsymbol{\varepsilon}}_t$ is a vector of mutually and serially independent exogenous error variables,

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

and $\tilde{\mathbf{\Gamma}}_0$ has unit diagonal. Rearranging terms such that $(\tilde{\mathbf{I}} - \tilde{\mathbf{\Gamma}}_0)\tilde{\mathbf{X}}_t$ is on the rhs ($\tilde{\mathbf{I}}$ being the identity matrix) should make clear that the "contemporaneous" causal relations are collected in $\tilde{\mathbf{\Gamma}}_0$. We assume $\tilde{\mathbf{\Gamma}}_0$ can be made lower triangular, i.e., that there is a *recursive causal ordering*. We limit our attention to linear and recursive data-generating processes in order to take advantage of existing theoretical results, and because linearity is typically assumed in macroeconometric practice; however, generalizing our approach to nonparametric and possibly non-recursive settings is an important direction for future research.

In the graphical causal modeling literature, systems of structural equations correspond to directed acyclic graphs (DAGs) that encode qualitative features of the equations, e.g., which variables are causally related and what conditional independence relationships are implied by the model. A DAG $\mathcal{G}$ is a pair $(\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is a set of vertices and $\mathbf{E}$ a set of directed edges ($\rightarrow$) connecting vertices subject to the restriction that there is no sequence of directed edges from any vertex to itself (no cycles). Vertices in $\mathbf{V}$ typically index some set of random variables. Specifically, corresponding to (1) is a dynamic DAG $\mathcal{G}$ with vertices corresponding to elements of $\{\tilde{\mathbf{X}}_t\}_{t\in\mathbb{N}}$, such that non-zero coefficients $(\tilde{\mathbf{\Gamma}}_s)_{ij} \neq 0$ $s \in (1, ..., p)$ and $(\tilde{\mathbf{I}} - \tilde{\mathbf{\Gamma}}_0)_{ij} \neq 0$ imply directed edges $V_{j,t-s} \rightarrow V_{i,t}$ and $V_{j,t} \rightarrow V_{i,t}$ (respectively) in $\mathcal{G}$. If $V_{j,t'} \rightarrow V_{i,t}$ in $\mathcal{G}$ we say $V_{j,t'}$ is a parent (direct cause) of $V_{i,t}$, and denote the set of parents of $V_{i,t}$ in $\mathcal{G}$ by $pa(V_{i,t}, \mathcal{G})$. Note that dynamic DAGs have an infinite number of vertices, since the SVAR model is defined for all $t \in \mathbb{N}$. However, in practice we only handle finite segments, since these models exhibit a repeating structure. An example is in Figure 1a. Restricting attention to only observed variables, the system may be represented by a dynamic MAG (maximal ancestral graph), which is a kind of mixed graph that includes both directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges [21, 26]. The latter correspond roughly to dependencies induced by latent common causes in the underlying "full" model. An example is in Figure 1b. See [15] for more details on dynamic DAGs, dynamic MAGs, and their causal interpretations.

Previous work on learning causal graphical models from time series data has largely ignored the possibility of relevant unmeasured processes and focused mostly on learning only the contemporaneous relations, i.e., zeros in the (marginal) matrix $\mathbf{\Gamma}_0$. Building on work initiated by [24], the authors of [2, 4, 9, 16, 5] have used graphical or related methods to learn contemporaneous causal relationships, having assumed that all relevant variables are measured. Typically, learning proceeds by estimating a reduced form vector autoregression (in some cases, in error-correction form) and executing a search procedure like PC [23] or LiNGaM [22] on the residuals, which corresponds to finding zeros in $\mathbf{\Gamma}_0$. Recently, [15] introduced methods for learning dynamic graphical models with unrestricted latent components. These algorithms, called SVAR-FCI and SVAR-GFCI, accomodate unmeasured variables by focusing on the dynamic MAG representation. More precisely, since SVAR-FCI and SVAR-GFCI use only conditional independence information to narrow down the range of causal models consistent with observed data, these algorithms may only recover a Markov equivalence class of dynamic MAGs, i.e., a set of models which all imply the same conditional independence relations. A Markov equivalence class of dynamic MAGs is represented by a dynamic PAG (partial ancestral graph). PAGs include $\circ\!\!\rightarrow$ and $\circ\!\!-\!\!\circ$ edges in addition to the possible edges in a MAG, where the $\circ$ mark represents uncertainty about the underlying orientation (i.e, it could be a "tail" or an "arrowhead").

SVAR-FCI is a constraint-based method, which uses sequential hypothesis tests of conditional independence to perform model selection. SVAR-GFCI is a hybrid score-based method, which uses greedy optimization of a model score to learn most of the connections in a graph, followed by some additional independence tests. Both algorithms use a variation of the orientation rules from the FCI algorithm [23, 27]. The greedy score-based approach relies on the availability of some model score satisfing three abstract properties: decomposability (that the score can be decomposed into a sum of "local" contributions), score-equivalence (that Markov equivalent graphs yield the same score), and consistency (that the score selects the true model in the limit). These properties imply an important property, called *local consistency*, that is sufficient to prove consistency of the search procedure, as we discuss below [3]. The BIC score is a popular choice that satisfies these properties, and is indeed used in conjunction with SVAR-GFCI in [15]. Our goal will be to devise a score which satisfies these properties in a nonstationary setting. Then, we can perform model selection with SVAR-GFCI, which we describe in detail in the supplementary material. First, we elaborate why stochastic trends pose a problem for traditional structure learning techniques.

2.2 Stochastic trends

Some nonstationary data is nonstationary because at least one variable in the system (perhaps a latent) exhibits a stochastic trend. When such a variable causally affects other variables, several variables may exhibit trending behavior and some subsets of the variables may be cointegrated. This poses a problem for using conditional independence-based methods of causal

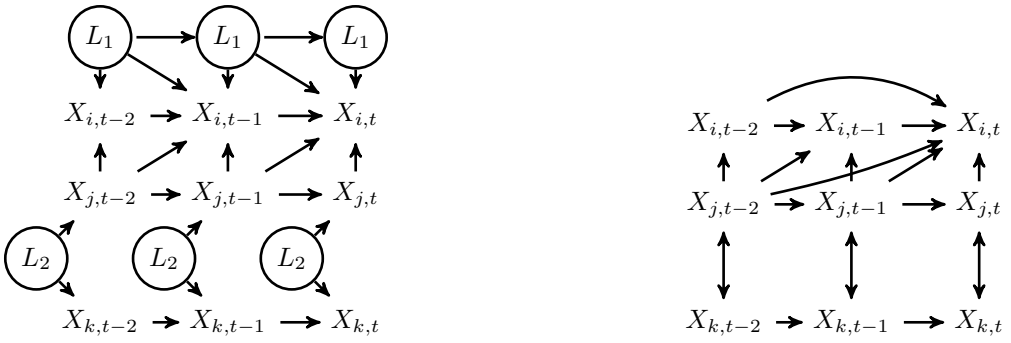

Figure 1: a) A dynamic DAG with latent processes. b) The implied dynamic MAG over the observed processes. Note that some latents can induce apparently "infinite-order" dependencies in the MAG; see [15] for an explanation.

structure learning. It is well-known that traditional tests of conditional independence or non-correlation cannot be straightforwardly applied to variables exhibiting stochastic trends, since such variables will appear correlated even when the processes are independent [25, 6, 7, 12]. Specifically, standard correlation tests rely on estimators of population-level correlation, such as the sample Pearson correlation coefficient, which in nonstationary settings fail to converge to any true value that measures correlation (or dependence) between the processes. However, in practice the observations will often be difference-stationary: when X_t exhibits a stochastic trend $\Delta X_t \equiv X_t - X_{t-1}$ may be stable and therefore stationary. That is, the variables will be I(1) (integrated of order one). One seemingly natural strategy to try is to use conditional independence tests among the differenced variables to search for a causal structure. The resultant model would be a causal graph over a transformed variable set, but presumably the model would bear some relationship to the original (not differenced, i.e., in levels) data-generating process.

Such a procedure would produce inaccurate results, because differencing can induce spurious correlations. To see this, consider a simple two-variable example.[1]

Assume the data-generating process is a linear, two-variable structural VAR with no contemporaneous causal connection:

$$\begin{aligned} X_t &= \alpha X_{t-1} + \beta Y_{t-1} + \varepsilon_t^X \\ Y_t &= \delta X_{t-1} + \gamma Y_{t-1} + \varepsilon_t^Y \end{aligned} \tag{2}$$

Say that the coefficients are chosen such that there is one unit root in the characteristic polynomial, ensuring stochastic nonstationarity. ε_t^X and ε_t^Y are mutually and serially independent Gaussian random variables. Note that X and Y mutually cause each other at one lag, but there is not contemporaneous causation between them. The differenced transformation of this model is simply obtained by carrying-through the difference operator:

$$\begin{aligned} \Delta X_t &= \alpha \Delta X_{t-1} + \beta \Delta Y_{t-1} + \Delta \varepsilon_t^X \\ \Delta Y_t &= \delta \Delta X_{t-1} + \gamma \Delta Y_{t-1} + \Delta \varepsilon_t^Y \end{aligned} \tag{3}$$

The structural coefficients relating differenced variables are just the same as those in the original data-generating process (2), so one may hope that a procedure which learns the qualitiative structure of (3) can be directly leveraged in service of estimating the coefficients in (2). (Parameter estimation, presumably, would be done in a separate step once the structure is known, using something like maximum likelihood techniques for nonstationary data.) Unfortunately, though it is not obvious from the written form of (3), the differenced variables ΔX_t and ΔY_t are (contemporaneously) dependent. To see this, consider a special case of the model with coefficients $\delta = \gamma = 0$ and $\alpha = 1$, so that X_t behaves like a random walk influenced by a stationary Y_t.

$$\begin{aligned} X_t &= X_{t-1} + \beta Y_{t-1} + \varepsilon_t^X \\ Y_t &= \varepsilon_t^Y \end{aligned} \tag{4}$$

From these equations one can derive:

$$\begin{aligned} \Delta X_t &= \beta \varepsilon_{t-1}^Y + \varepsilon_t^X \\ \Delta Y_t &= \varepsilon_t^Y - \varepsilon_{t-1}^Y \end{aligned} \tag{5}$$

Notice how the expressions for ΔX_t and ΔY_t in terms of purely stationary components both include ε_{t-1}^Y, so they will be correlated. Another way of putting this is that differencing has induced correlation among the error terms in system (4): $\Delta \varepsilon_t^X$ and $\Delta \varepsilon_t^Y$ are both serially correlated ($\Delta \varepsilon_t^X$ is correlated with $\Delta \varepsilon_{t-1}^X$ because they overlap) and mutually correlated.

Taking all the dependencies among the errors into account, the graphical representation corresponding to

[1] The example and problem were suggested by Kevin Hoover, personal communication.

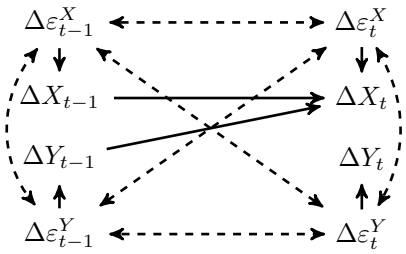

Figure 2: The graphical representation of the differenced version of system (4). Here dashed double-headed edges represent correlations among the errors induced by differencing, and the solid directed edges correspond to true structural connections among levels.

the differenced version of system (4) (including the error variables) is in Figure 2. While this model shares some features with the underlying data-generating process (4), the correlated errors ensure that the output of a typical causal search algorithm applied to these differenced variables would be quite misleading. This simple example also highlights a problem with a suggestion due to Moneta et al. [16, p. 103]. In that paper, the authors propose a procedure which involves causal search on the estimated residuals from a VAR model, along the lines of work discussed earlier. Specifically, where the researcher has I(1) variables but detects no cointegration, Moneta et al. prescribe estimating the VAR in first differences, and performing tests of vanishing partial correlation on the residuals as input into a search algorithm such as PC. This would not produce correct causal conclusions, since patterns of partial correlation among the residuals from differences do not correspond to patterns of partial correlation among the original variables in levels.

Since structure learning applied to first differences will fail to reproduce the structural relationships in the data-generating process (in levels), we instead proceed by making use of likelihood calculations in a reparameterization of the data-generating process.

3 An error-correction approach

In applied econometric work, analysis of nonstationary data which is cointegrated typically proceeds by re-specifying the data-generating process in VECM ("vector error-correction model") form. We begin by assuming the data-generating process is (1), and allow that the system is not stable (the matrix coefficients fail to satisfy the usual eigenvalue condition), but assume that all variables become stable upon differencing [14]. Thus, the process admits nonstationary solutions, including cointegration relations. The VECM paramterization of (1) is written in terms of both differences and levels, as follows:

$$\begin{aligned}\boldsymbol{\Delta}\tilde{\mathbf{X}}_t =&(\tilde{\mathbf{I}} - \tilde{\boldsymbol{\Gamma}}_{\mathbf{0}})\boldsymbol{\Delta}\tilde{\mathbf{X}}_t + \tilde{\boldsymbol{\Pi}}\tilde{\mathbf{X}}_{t-1} \\ &+ \tilde{\mathbf{B}}_1\boldsymbol{\Delta}\tilde{\mathbf{X}}_{t-1} + ... + \tilde{\mathbf{B}}_{p-1}\boldsymbol{\Delta}\tilde{\mathbf{X}}_{t-p-1} + \tilde{\boldsymbol{\varepsilon}}_t\end{aligned} \tag{6}$$

The parameters are related to those in (1) by $\tilde{\boldsymbol{\Gamma}}_1 = \tilde{\boldsymbol{\Pi}} + \tilde{\boldsymbol{\Gamma}}_0 + \tilde{\mathbf{B}}_1$, $\tilde{\boldsymbol{\Gamma}}_i = \tilde{\mathbf{B}}_i - \tilde{\mathbf{B}}_{i-1} (i = 2, ..., p-1)$, and $\tilde{\boldsymbol{\Gamma}}_p = -\tilde{\mathbf{B}}_{p-1}$. Notice that despite the appearance of differences in the VECM form, the errors are preserved, i.e., they are the same errors as in the original SVAR in levels. Also, we preserve the contemporaneous relationships in $\tilde{\boldsymbol{\Gamma}}_0$, so this model may sometimes be referred to as the *structural* VECM. The cointegration typical of stochastic trend models manifests in the reduced rank of the coefficient matrix $\boldsymbol{\Pi}$ in the marginal model. Cointegration can be attributed to some number of "driving trends," i.e., possibly latent common causes of measured variables which themselves behave as random walks, and which are thus simultaneously responsible for the nonstationarity of their effects and for the cointegrated behavior [8]. In this case, the driving trends may be included in $\mathbf{L}_t$, though we do not place any specific restrictions on $\mathbf{L}_t$, except that the components are I(1). In contrast to typical cointegration analysis, our aim is not to estimate the number of cointegrating relationships or precisely which linear combinations of levels are stationary, but rather to learn the structural relationships in (1) among the observed variables. Facts about the cointegrating relations can subsequently be inferred from the estimated model, under some assumptions.

The VECM parameterization (6) is useful in a score-based approach because we can straightforwardly consistently estimate the parameters and thus calculate maximum likelihood for a candidate model. If we can calculate the maximum likelihood, then we can calculate the model's BIC score. So, a greedy score-based procedure is feasible: for every candidate causal edge, we derive the corresponding structural VECM parameterization, estimate its parameters, calculate the BIC score difference, and add or remove the candidate edge depending on whether the score improves. In its first phase, the SVAR-GFI algorithm searches for the best-scoring DAG model, even though no DAG perfectly describes the distribution due to the presence of latent variables. In the subsequent stages of the algorithm, SVAR-GFCI performs conditional independence tests to remove some edges and then propagates orientation rules to transform the result into a valid PAG (see the supplement). For these subsequent conditional independence tests, we can use the BIC score again. Since the score-based phase of the algorithm considers candidate DAG models, our discussion of scores in the next section is concerned with scoring DAGs.

3.1 The BIC score for integrated data and as independence test

Consider two candidate dynamic DAG models, $\mathcal{G}_1$ and $\mathcal{G}_2$, corresponding to fully oberved SVARs. Let the pair of vertices $(X_{i,t}, X_{j,s})$ be called *homologous* to pair $(X_{m,a}, X_{n,b})$ if $m = i$, $n = j$, and $t - s = a - b$. In our structure learning procedure we consider insertions and deletions of homologous sets of edges since in the SVAR parameterization the coeffiecients for $X_{i,t-1} \to X_{j,t-1}$ and $X_{i,t} \to X_{j,t}$ (etc.) are the same parameter. Without loss of generality assume $\mathcal{G}_2$ has one additional homologous set of edges but is otherwise the same as $\mathcal{G}_1$. For each model we can derive the corresponding VECM form as in (6). The BIC score for a model $\mathcal{G}_i$ is

$$\mathrm{BIC}_i = \log p(\mathbf{D}|\hat{\mathbf{\Theta}}, \mathcal{G}_i) - \frac{d}{2}\log T \tag{7}$$

where $\mathbf{D}$ is the observed data, $\hat{\mathbf{\Theta}}$ is the vector of parameters that maximize the likelihood for model $\mathcal{G}_i$, T is the sample size, and d is the number of estimated parameters. Typically for a model corresponding to an I(1) process, the log-likelihood term would be calculated for the corresponding VECM parameterization using reduced rank regression, conditional on some "known" rank reduction determined from a prior sequence of tests [11, 14, p. 294-5]. In our case, we do not impose any rank restrictions. This is important for the greedy search strategy we employ which considers only single-edge changes to a candidate model, thus requiring only calculations of "local" score differences [3]. Alternative procedures may consider incorporating rank reduction information, but at the cost of inducing cross-equation parameter dependence and thus forgoing the computational scaleability of "local" score-based structure learning.

In our case, we may compare two candidate models $\mathcal{G}_1$ and $\mathcal{G}_2$ that differ by one homologous set of edges by estimating the corresponding parameters in the VECM form, and subsequently the BIC score difference (or rather, the "local" contribution to the score difference attributed to the single-parameter change). The difference in dimension is always one, since we only consider single-parameter differences in the course of greedy search.

Recall the definition of a *locally consistent* score [3]:

Definition 3.1. *Let $\mathcal{G}$ be any DAG, and let $\mathcal{G}'$ be the DAG that results from adding the edge $X_i \to X_j$. A scoring criterion $S(\mathcal{G}, \mathbf{D})$ is locally consistent if the following two properties hold:*

1. If $X_i \not\perp\!\!\!\perp X_j | pa(X_j, \mathcal{G})$ then $\lim_{T\to\infty} \mathbb{P}(S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})) = 1$

2. If $X_i \perp\!\!\!\perp X_j | pa(X_j, \mathcal{G})$ then $\lim_{T\to\infty} \mathbb{P}(S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})) = 1$

[3, Lemma 7] proves the local consistency of the BIC score in the i.i.d. sampling and Gaussian setting. From local consistency of the BIC score one may derive the following observation:[2]

$$X_i \perp\!\!\!\perp X_j | \mathbf{Z} \text{ iff } \mathrm{BIC}_1 - \mathrm{BIC}_2 > 0 \tag{8}$$

in the limit as $T \to \infty$ where $\mathcal{G}_2$ and $\mathcal{G}_1$ are identical except for the additional edge $X_i \to X_j$ in $\mathcal{G}_2$, and $\mathbf{Z}$ is identified with the set of parents of X_j in $\mathcal{G}_1$. In the Gaussian setting, the maximum log-likelihood for X_j is

$$-T\log(2\pi) - \frac{T}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{t=1}^{T}(\hat{\varepsilon}_t)^2$$

where where $\hat{\sigma}^2$ and $\hat{\varepsilon}_t$ are the ML estimates of the variance and residuals. Let $RSS_i = \sum_t^T (\hat{\varepsilon}_t)^2$ be the residual sum of squares estimated from each specification ($i \in \{1, 2\}$), and consider estimating the variance for both models with the unbiased estimator for the larger model, as is commonly done: $\hat{\sigma}^2 = \frac{RSS_2}{T-(d+1)}$. Then we have

$$\mathrm{BIC}_1 - \mathrm{BIC}_2 > 0 \text{ iff}$$
$$\frac{1}{\hat{\sigma}^2}(RSS_1 - RSS_2) < \log T$$

The term on the left hand side is distributed $F_{1,T-d-1}$. (Without using the $\hat{\sigma}^2 = \frac{RSS_2}{T-(d+1)}$ estimator for both variances, one may derive the same test statistic but only after a Taylor expansion.) Thus, using the BIC score difference is equivalent to a hypothesis test with critical value varying as a function of sample size. Hypothesis testing with critical value depending on sample size (approximately) manifests the Bayesian alternative to classical hypothesis testing at fixed level [10].

To see how this works practically in the I(1) setting, it may help to consider a simple example with 3 processes (all assumed to be observed), and $p = 2$. First note:

$$\begin{aligned}
\mathbf{\Gamma}_0\mathbf{X}_t &= \mathbf{\Gamma}_1\mathbf{X}_{t-1} + \mathbf{\Gamma}_2\mathbf{X}_{t-2} + \boldsymbol{\varepsilon}_t \\
\mathbf{\Gamma}_0\mathbf{\Delta}\mathbf{X}_t &= (\mathbf{\Gamma}_1 - \mathbf{\Gamma}_0 + \mathbf{\Gamma}_2)\mathbf{X}_{t-1} - \mathbf{\Gamma}_2\mathbf{\Delta}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t \\
\mathbf{\Delta}\mathbf{X}_t &= (\mathbf{I} - \mathbf{\Gamma}_0)\mathbf{\Delta}\mathbf{X}_t + (\mathbf{\Gamma}_1 - \mathbf{\Gamma}_0 + \mathbf{\Gamma}_2)\mathbf{X}_{t-1} \\
&\quad - \mathbf{\Gamma}_2\mathbf{\Delta}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t
\end{aligned}$$

Consider a data-generating process for $\mathbf{X}_t = (X_{1,t}, X_{2,t}, X_{3,t})'$ where

$$\mathbf{\Gamma}_0 = \begin{bmatrix} 1 & -\gamma^0_{12} & 0 \\ 0 & 1 & -\gamma^0_{23} \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{\Gamma}_1 = \begin{bmatrix} \gamma^1_{11} & \gamma^1_{12} & 0 \\ \gamma^1_{21} & \gamma^1_{22} & 0 \\ 0 & 0 & \gamma^1_{33} \end{bmatrix}$$

[2] The idea of using the BIC score difference as a general independence test was introduced in [20]. This holds whenever the BIC score is locally consistent. Relatedly, [17] connect BIC score differences to vanishing partial correlations in i.i.d. multivariate Gaussian models. See also [1] for prior work on the relationship between the BIC score and F-tests.

$$\mathbf{\Gamma}_2 = \begin{bmatrix} 0 & \gamma_{12}^2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Assume the true parameter values are such that the characteristic polynomial has a unit root, but differences are stable.[3] Write X_t, Y_t, Z_t for $X_{1,t}, X_{2,t}, X_{3,t}$ for readability in what follows. In learning the structure of such a model, we compare changes to BIC scores from inserting or deleting edges, essentially testing whether the elements γ_{ij}^s are zero, conditional on other edges being present. For example: at some stage of the search procedure the edge $Y_t \to X_t$ is considered, given that the parent set of X_t has been determined (at an earlier stage) to include true parents X_{t-1}, Y_{t-1}, and Y_{t-2} (i.e., $\gamma_{11}^1, \gamma_{12}^1$, and γ_{12}^2 have been judged to be non-zero). That is, the algorithm needs to decide whether $\gamma_{12}^0 = 0$, equivalent to deciding whether $Y_t \perp\!\!\!\perp X_t | \{X_{t-1}, Y_{t-1}, Y_{t-2}\}$. Regressing X_t on Y_t and the other parents would produce spurious results since X_t is nonstationary. Instead, we calculate the local difference in BIC scores by comparing RSS from two regressions with differenced dependent variables, the latter imposing $\gamma_{12}^0 = 0$:

$$\begin{aligned} &\Delta X_t \sim \Delta Y_t + X_{t-1} + Y_{t-1} + \Delta Y_{t-1} \\ &\quad \text{versus} \\ &\Delta X_t \sim X_{t-1} + Y_{t-1} + \Delta Y_{t-1} \end{aligned}$$

So the procedure here is straightforward. The score difference is just calculated from the RSS difference for these two VECM regressions (and $\log T$). The analagous procedure would determine if any γ_{ij}^2 is zero, e.g., for γ_{12}^2 just add or drop ΔY_{t-1} from the corresponding regressions. Determining the zeros in $\mathbf{\Gamma}_1$ requires somewhat more care, since the matrix of coefficients on $\mathbf{X}_{t-1}$ in the VECM parameterization is a mixture of the fundamental structural matrices $(\mathbf{\Gamma}_1 - \mathbf{\Gamma}_0 + \mathbf{\Gamma}_2)$. To determine, for example, whether $\gamma_{12}^1 = 0$, we cannot simply add/drop Y_{t-1} from the VECM regression of ΔX_t on $\Delta Y_t, X_{t-1}$, and ΔY_{t-1} since the coefficient on Y_{t-1} in the VECM parameterization is $\gamma_{12}^1 + \gamma_{12}^0 + \gamma_{12}^2$. We want to check whether $\gamma_{12}^1 = 0$ in the equation for ΔX_t, which is

$$\begin{aligned} \Delta X_t = \gamma_{12}^0 \Delta Y_t + (\gamma_{11}^1 - 1) X_{t-1} \\ + (\gamma_{12}^1 + \gamma_{12}^0 + \gamma_{12}^2) Y_{t-1} - \gamma_{12}^2 \Delta Y_{t-1} + \varepsilon_{1,t}. \end{aligned}$$

[3]Note that we assume here and throughout only that the parameters in the true model have this property, not necessarily every submodel or supermodel we consider in the course of search. This is important because necessarily in the course of our procedure we consider setting various parameters in the $\mathbf{\Gamma}$ matrices to zero, and not all such models will have the same properties vis-à-vis unit roots.

This can be rearranged to

$$\begin{aligned} \Delta X_t &= \gamma_{12}^0 (\Delta Y_t + Y_{t-1}) + (\gamma_{11}^1 - 1) X_{t-1} \\ &\quad + \gamma_{12}^1 Y_{t-1} + \gamma_{12}^2 (Y_{t-1} - \Delta Y_{t-1}) + \varepsilon_{1,t} \\ &= \gamma_{12}^0 Y_t + (\gamma_{11}^1 - 1) X_{t-1} + \gamma_{12}^1 Y_{t-1} \\ &\quad + \gamma_{12}^2 Y_{t-2} + \varepsilon_{1,t} \end{aligned}$$

So we compare:

$$\begin{aligned} &\Delta X_t \sim Y_t + X_{t-1} + Y_{t-1} + Y_{t-2} \\ &\quad \text{versus} \\ &\Delta X_t \sim Y_t + X_{t-1} + Y_{t-2} \end{aligned}$$

The difference in the RSS from these two specifications gives us the difference in BIC scores from inserting or removing $Y_{t-1} \to X_t$ given the other parents of X_t. In this case the algorithm is deciding whether $Y_{t-1} \perp\!\!\!\perp X_t | \{Y_t, X_{t-1}, Y_{t-2}\}$.

We can generalize these observations into a rule for calculating score comparisons for insertion and deletion of edges. Let the $\widetilde{pa}(X_{i,t}, \mathcal{G})$ be the transformed parent set of $X_{i,t}$ in VECM form, constructed as follows $(\forall j)$:

$$\begin{aligned} X_{j,t} \in pa(X_{i,t}, \mathcal{G}) &\Rightarrow \{\Delta X_{j,t}, X_{j,t-1}\} \in \widetilde{pa}(X_{i,t}, \mathcal{G}); \\ X_{j,t-1} \in pa(X_{i,t}, \mathcal{G}) &\Rightarrow X_{j,t-1} \in \widetilde{pa}(X_{i,t}, \mathcal{G}); \\ X_{j,t-2} \in pa(X_{i,t}, \mathcal{G}) &\Rightarrow \{\Delta X_{j,t-1}, X_{j,t-1}\} \in \widetilde{pa}(X_{i,t}, \mathcal{G}). \end{aligned}$$

This is the case for $p = 2$, but the principle is that $\widetilde{pa}(X_{i,t}, \mathcal{G})$ contains all VECM variables (i.e., differences and lagged levels) which have non-zero coefficients in the VECM form of the underlying SVAR. Then the estimated residuals $\hat{\varepsilon}_{i,t}$ from the least squares regression $\Delta X_{i,t} \sim \widetilde{pa}(X_{i,t}, \mathcal{G})$ are consistent estimates of the underlying error terms $\varepsilon_{i,t}$.

We use $S(\Delta X_{i,t} | \widetilde{pa}(X_{i,t}, \mathcal{G}), \mathbf{D})$ to denote local contributions to the model score. The local contribution to the likelihood is derived from the residuals of the least squares regression of $\Delta X_{i,t}$ on $\widetilde{pa}(X_{i,t}, \mathcal{G})$. Consequently, when considering the edge addition $X_{j,t-s} \to X_{i,t}$ $(s \geq 0)$ in graph $\mathcal{G}$, the BIC score difference is given by:

$$\begin{aligned} &S(\Delta X_{i,t} | \widetilde{pa}(X_{i,t}, \mathcal{G}), \mathbf{D}, \gamma_{ij}^s = 0) \\ -&S(\Delta X_{i,t} | \widetilde{pa}(X_{i,t}, \mathcal{G}), \mathbf{D}, \gamma_{ij}^s \neq 0) \end{aligned}$$

where $\gamma_{ij}^s = 0$ and $\gamma_{ij}^s \neq 0$ indicate that the coefficient corresponding to $X_{j,t-s} \to X_{i,t}$ is constrained to be zero and non-zero respectively in the corresponding regressions. There are perhaps multiple ways of implementing these score-difference calculations; in our software, we implement the calculation by rearranging terms such that we only need to drop or add regressors from the appropriate regressions to impose $\gamma_{ij}^s = 0$ or $\gamma_{ij}^s \neq 0$, along the lines described in the three-variable example above.

3.2 Consistency of the procedure

Here we spell out the assumptions required for consistency a bit more precisely. We make the following assumptions:

A1 The data-generating process is $\tilde{\mathbf{\Gamma}}_0\tilde{\mathbf{X}}_t = \tilde{\mathbf{\Gamma}}_1\tilde{\mathbf{X}}_{t-1} + ... + \tilde{\mathbf{\Gamma}}_p\tilde{\mathbf{X}}_{t-p} + \tilde{\boldsymbol{\varepsilon}}_t \ \forall t \in \mathbb{N}$, conditional on the initial values $(\tilde{\mathbf{X}}_{-p+1}, ..., \tilde{\mathbf{X}}_0)'$ with $\tilde{\mathbf{X}}_t = (\mathbf{L}_t^{'}, \mathbf{X}_t^{'})'$, $\tilde{\mathbf{\Sigma}} = \mathbb{E}[\tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t']$ diagonal, and $\tilde{\boldsymbol{\varepsilon}}_t \sim_{iid} N(0, \tilde{\mathbf{\Sigma}})$. $\tilde{\mathbf{\Gamma}}_0$ has unit diagonal and can be made lower triangular.

A2 The stochastic process $\{\tilde{\mathbf{X}}_t\}_{t\in\mathbb{N}}$ is I(1), i.e., $\{\boldsymbol{\Delta}\tilde{\mathbf{X}}_t\}_{t\in\mathbb{N}}$ is stable.

A3 Let $p_t(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_{t-1}, ..., \tilde{\mathbf{X}}_{t-p})$ be the joint density at time t (conditional on initial values). $p_t(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_{t-1}, ..., \tilde{\mathbf{X}}_{t-p})$ is Markov and faithful to a dynamic DAG $\mathcal{G}$ for all t.

A3 allows that the joint distribution is the not same at all times (nonstationarity), but assumes it always factorizes according to the same graph, and that elements of the process are conditionally independent iff they are d-separated in that graph (the faithfulness assumption). See [23] for relevant background on faithfulness. The Gaussianity assumption in A1 guarantees that Markov equivalent models will have the same likelihood.

We define the VECM BIC score for a candidate DAG model as:

$$\frac{T}{2}\log|\hat{\mathbf{\Sigma}}| - \frac{d}{2}\log T \tag{9}$$

where $\hat{\mathbf{\Sigma}}$ is derived from the least squares estimated residuals of the corresponding VECM over observed variables, d is the number of free parameters, and T is the sample size. Proofs of the following propositions can be found in the supplement.

Proposition 1. *The VECM BIC score is locally consistent under assumptions A1 and A2.*

As a corollary of Proposition 1, we have an asympotically consistent test of conditional independence for I(1) data, using the result (8): call this the BIC CI test. With a locally consistent score and consistent independence test, the SVAR-GFCI algorithm will asymptotically select the true PAG [15].

Proposition 2. *Assume the stochastic process $\{\tilde{\mathbf{X}}_t\}_{t\in\mathbb{N}}$, where $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$, satisfies A1-A3. Let $\mathcal{M}$ be the MAG implied by $\mathcal{G}$ over $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$ and PAG $\mathcal{P}$ the equivalence class of $\mathcal{M}$. Given T observations of the marginal subprocess $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$, the SVAR-GFCI algorithm with the VECM BIC score and BIC CI test is a consistent estimator of $\mathcal{P}$.*

Furthermore, the BIC CI test makes entirely constraint-based search possible, and so in fact we may claim the same property for SVAR-FCI when equipped with this test. However, we limit our attention to SVAR-GFCI here.

4 Simulation experiments

To explore the prospects of structure learning from I(1) data with the VECM BIC score, we carried out a simulation study on a small I(1) model. The data-generating process is a dynamic DAG (shown in Figure 1 in the supplementary material) that includes 5 measured processes $X_{1,t}, ..., X_{5,t}$ as well as two exogenous latent confounding processes $L_{1,t}, L_{2,t}$. The latent processes are pure random walks, i.e., $L_{i,t} = L_{i,t-1} + \varepsilon_{i,t}$, which induce stochastic nonstationarity among the observed variables. All parameters in the observed processes are chosen such that they do not create their "own" nonstationary behavior – specifically, all edges connected to measured variables are parameterized with linear coefficients set to 0.55; the $L_{i,t-1} \to L_{i,t}$ edges carry coefficients of 1.0. All error terms are distributed $N(0, 1)$. The features of this model – specifically the purely exogenous latent trends which are random walks, affecting some combinations of variables but not all variables, the relatively sparse structure – are motivated by common features of models in the cointegrated VAR literature [13, 8].

We generated data from this model at varying sample size and learned a dynamic PAG from each data set using SVAR-GFCI with an implementation of the VECM BIC score. Average adjacency precision and recall statistics are reported in Figure 3, with 100 runs at each sample size (new initial values were sampled from $N(0, 1)$ for each run). We also include, for comparison, performance results obtained with SVAR-GFCI using the usual (unadjusted for nonstationarity) BIC score, which does quite poorly as expected and in fact increasingly worse with sample size.

The precision and recall results support our claim that it is possible to perform structure learning from I(1) series with a slightly modified score-based procedure. We see that precision is quite high consistently, and recall is lower, though increasing with sample size. Note that high precision means that there are almost no false positives (we see an average of about one false positive edge for these settings), in contrast with what may be expected from the phenomenon of "spurious regression," where nonstationary levels regressed on nonstationary levels frequently produce false judgements of dependence between independently generated processes. SVAR-GFCI using the unadjusted BIC score for stationary data does poorly, much worse on both precision and recall (except for small sample size precision of arrowheads, where the procedures perform

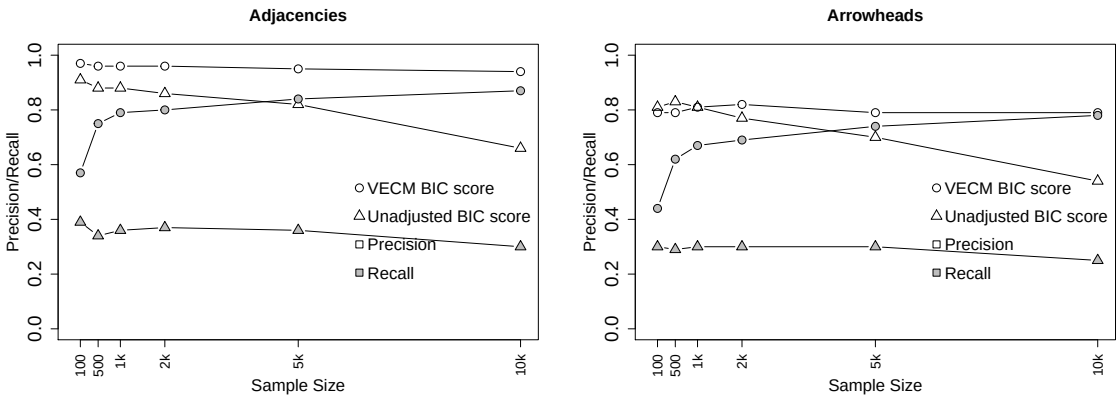

Figure 3: Precision (white) and recall (grey) results from the nonstationary simulation study.

about the same).[4]

5 Danish macroeconomic data

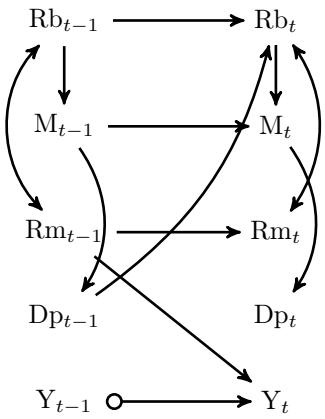

Figure 4: Results on the Danish data. *Rb* = interest rate on govt bonds, *Rm* = own interest rate on money stock, M = log real money, *Dp* = inflation, Y = log aggregate real income.

We apply our method to a data set of several I(1) macroeconomic quantities – inflation rates, money stock, aggregate income, and two interest rates – observed quarterly in Denmark between 1973 and 2003:1 ($T = 121$). The data has been extensively examined in [13] as well as other sources, and previously been used by the Central Bank of Denmark for policy analysis. In [13], the author considers multiple different VECM model specifications, some of which are inspired by theoretical hypotheses (e.g., how money drives inflation, equilibrium in the money market, etc.) and some by empirical regularities, such as persistent cointegration relationships found in the data. In our data-driven model specification discovered by SVAR-GFCI (Figure 4), we find some noteworthy features: namely, that the two interest rates are linked by a common latent mechanism (as arguably predicted by equilibrium in the money market), and that (lagged) inflation does influence money indirectly through interest rates, but also that the causal influence propagates back to inflation, i.e., there is a feedback loop.

6 Conclusion

In this paper we extended previously introduced structure learning techniques to a nonstationary setting. Stochastic trends, at least in the common case where variables are I(1), are handled by replacing the usual BIC score with the VECM BIC score; this takes advantage of a re-parameterization of the data-generating process in error-correction form, and then estimates those parameters by maximum likelihood. Throughout, we have assumed that the underlying data-generating process is an SVAR in levels (with latent processes), and the goal has been to recover as much information as possible about the causal relations among observed variables. Though the score-based procedure permits cointegration relationships among some of the variables, it does not make any use of such information even if it is available. Thus, possible avenues for future research include developing techniques which explicitly incorporate cointegration information, or which shift the focus to identifying features of the latent structure, e.g., determining which latent processes are causes of which measured processes, or perhaps what are the causal relations among the latent variables themselves.

[4] We find that at small samples sizes, many edges are falsely removed (low recall on adjacencies) and so there are many false unshielded triples oriented as colliders in subsequent stages of the algorithm. With too few edges and too many colliders, the output has high arrowhead precision for the "wrong reasons."

Acknowledgements

The authors would like to thank David Danks, Clark Glymour, Kevin Hoover, Søren Johansen, and Joseph Ramsey. A special thanks is due to Katarina Juselius for providing the Danish economic data. Research reported in this publication was partially supported by NIH grant U54HG008540.

References

[1] Erdal Atukeren. The relationship between the F-test and the Schwarz criterion: implications for Granger-causality tests. *Economics Bulletin*, 30(1):494–499, 2010.

[2] David A Bessler and Seongpyo Lee. Money and prices: US data 1869–1914 (a study with directed graphs). *Empirical Economics*, 27(3):427–446, 2002.

[3] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[4] Selva Demiralp and Kevin D Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65(s1):745–767, 2003.

[5] Selva Demiralp, Kevin D Hoover, and Stephen J Perez. Still puzzling: evaluating the price puzzle in an empirically identified structural vector autoregression. *Empirical Economics*, 46(2):701–731, 2014.

[6] Kevin D Hoover. Nonstationary time series, cointegration, and the principle of the common cause. *The British Journal for the Philosophy of Science*, 54(4):527–551, 2003.

[7] Kevin D Hoover. Probability and structure in econometric models. In *Proceedings of the 13th International Congress of Logic, Methodology and Philosophy of Science*, pages 497–513, 2009.

[8] Kevin D Hoover. Long-run causal order: A preliminary investigation. Technical report, Economic Research Initiatives at Duke (ERID) Working Paper, 2018.

[9] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

[10] Harold Jeffreys. *The theory of probability*. Clarendon Press, 1939.

[11] Søren Johansen. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254, 1988.

[12] Søren Johansen. The analysis of nonstationary time series using regression, correlation and cointegration. *Contemporary Economics*, 6(2):40–57, 2012.

[13] Katarina Juselius. *The cointegrated VAR model: methodology and applications*. Oxford University Press, 2006.

[14] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[15] Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 23–47, 2018.

[16] Alessio Moneta, Nadine Chlaß, Doris Entner, and Patrik Hoyer. Causal search in structural vector autoregressive models. In *Proceedings of the NIPS Mini-symposium on Causality in Time Series*, pages 95–114, 2011.

[17] Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46(6A):3151–3183, 2018.

[18] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[19] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

[20] Joseph D Ramsey. Improving accuracy and scalability of the PC algorithm by maximizing p-value. *arXiv preprint arXiv:1610.00378*, 2016.

[21] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

[22] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[23] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

[24] Norman R Swanson and Clive WJ Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92(437):357–367, 1997.

[25] G Udny Yule. Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89(1):1–63, 1926.

[26] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

[27] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008.