Supplementary Materials for: Learning the Structure of a Nonstationary Vector Autoregression

Daniel Malinsky and Peter Spirtes
Johns Hopkins University and Carnegie Mellon University

1 Description of SVAR-GFCI

We reproduce several definitions and pseudcode from [4] to describe the SVAR-GFCI algorithm. The algorithm calls SVAR-GES as a subroutine, which is a modfied version of the GES algorithm [3]. As described in [4], SVAR-GES simply modifies the steps of GES to enforce the time order and repeating structure of the presumed SVAR data-generating process.

Definition 1.1. *Two variables are* adjacent *if there is some edge between them. Denote the set of adjacencies of X by $adj(X, \mathcal{G})$. A* path *is a sequence of distinct adjacent vertices containing at least two vertices, e.g.,* $\langle X_i, X_{i+1}, ..., X_{i+n}\rangle$. *A path is a* directed path from X_i to X_n *if for all* $m \in \{1, ..., n\}$ *the edge* $X_{i+m-1} \to X_{i+m}$ *occurs.*

Definition 1.2. *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* $t - s = a - b$. $hom(X_{i,t}, X_{j,s}, \mathcal{G})$ *denotes the set of vertex pairs homologous to* $(X_{i,t}, X_{j,s})$ *in graph* $\mathcal{G}$.

Definition 1.3. *Given a path π in a graph $\mathcal{G}$, a non-endpoint vertex X_j on π is called a* collider *if the two edges incident to X_j are both into X_j, i.e., have arrowheads at X_j ($*\!\!\to X_j \leftarrow\!\!*$). (Note the $*$ mark is used to represent any possible endpoint.) A* v-structure *is a triple $\langle X_i, X_j, X_k\rangle$ such that $X_i *\!\!\to X_j \leftarrow\!\!* X_k$ and X_i and X_k are not adjacent.*

Definition 1.4. *Let $X \in pds(X_i, X_j, \mathcal{G})$ if and only if $X \neq X_i$, $X \neq X_j$, and there is a path π between X_i and X in $\mathcal{G}$ such that for every subpath $\langle X_m, X_l, X_h\rangle$ of π either X_l is a collider on the subpath in $\mathcal{G}$ or $\langle X_m, X_l, X_h\rangle$ is a triangle in $\mathcal{G}$. A* triangle *is a triple $\langle X_m, X_l, X_h\rangle$ where each pair of vertices is adjacent.*

Definition 1.5. *Let* $adj_t(X_{i,t}, \mathcal{G}) = \{X_{j,s} : X_{j,s} \in adj(X_{i,t}, \mathcal{G}), s \leq t\}$ *and* $pds_t(X_{i,t}, X_{k,u}, \mathcal{G}) = \{X_{j,s} : X_{j,s} \in pds(X_{i,t}, X_{k,u}, \mathcal{G}), s \leq \max(t, u)\}$.

Note that SVAR-GFCI takes as input a score and an independence test. Typically, independence tests have a user-specified rejection threshold α. The BIC CI test as we defined it has no user-specified threshold, but instead the level of the test naturally varies with sample size in an approximately Bayesian way (as mentioned in the main text). So, α should be considered an optional input below. Of course, the score used with SVAR-GFCI in the present context is the VECM BIC score.

Algorithm 1.1: SVAR-GFCI(Score, Test, α)

Input: Data on variables $\mathbf{X}_t, ..., \mathbf{X}_{t-p} = \{X_{1,t}, ..., X_{k,t}, ..., X_{1,t-p}, ..., X_{k,t-p}\}$
Output: Dynamic PAG segment $\mathcal{P}$
1. $\mathcal{G} \leftarrow$ SVAR-GES(Score)
2. Form the graph $\mathcal{P}$ on vertex set $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$ with adjacencies in $\mathcal{G}$ and $\circ\!\!-\!\!\circ$ edges.
3. $n \leftarrow 0$
4. **repeat**
5. **for all** pairs of adjacent vertices $(X_{i,t}, X_{j,s})$ s.t. $|adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}| \geq n$ and subsets $\mathbf{S} \subset adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}$ s.t. $|\mathbf{S}| = n$
6. **if** $X_{i,t} \perp\!\!\!\perp X_{j,s} | \mathbf{S}$ according to (Test, α)

$$\textbf{then} \begin{cases} \text{Delete edge } X_{i,t} \circ\!\!-\!\!\circ X_{j,s} \text{ from } \mathcal{P}. \\ \text{Delete edge } X_{m,a} \circ\!\!-\!\!\circ X_{n,b}\ \forall (X_{m,a}, X_{n,b}) \in \textit{hom}(X_{i,t}, X_{j,s}, \mathcal{P}). \\ \text{Let } \textit{sepset}(X_{i,t}, X_{j,s}) = \textit{sepset}(X_{i,t}, X_{j,s}) = \mathbf{S}. \end{cases}$$

7. **end**
8. $n \leftarrow n + 1$
9. **until** for each pair of adjacent vertices $(X_{i,t}, X_{j,s})$, $|adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}| < n$.
10. **for all** adjacent vertices $(X_{i,t}, X_{j,s})$ orient $X_{i,t} *\!\!\rightarrow X_{j,s}$ iff $s > t$.
11. **for all** triples $(X_{i,t}, X_{k,r}, X_{j,s})$ s.t. $X_{i,t} \in adj_t(X_{k,r}, \mathcal{P})$ and $X_{j,s} \in adj_t(X_{k,r}, \mathcal{P})$ but $X_{i,t} \notin adj_t(X_{j,s}, \mathcal{P})$, orient $X_{i,t} *\!\!\rightarrow X_{k,r} \leftarrow\!\!* X_{j,s}$ iff $(X_{i,t}, X_{k,r}, X_{j,s})$ is a v-structure in $\mathcal{G}$, or it is a triangle in $\mathcal{G}$ and $X_{k,r} \notin \textit{sepset}(X_{i,t}, X_{j,s})$; then also orient $X_{m,a} *\!\!\rightarrow X_{o,c} \leftarrow\!\!* X_{n,b}$ $\forall (X_{m,a}, X_{o,c}) \in \textit{hom}(X_{i,t}, X_{k,r}, \mathcal{P})$ and $\forall (X_{n,b}, X_{o,c}) \in \textit{hom}(X_{j,t}, X_{k,r}, \mathcal{P})$
12. **for all** pairs $(X_{i,t}, X_{j,s})$ adjacent in $\mathcal{P}$ **if** $\exists \mathbf{S}$ s.t. $\mathbf{S} \in pds_t(X_{i,t}, X_{j,s}, \mathcal{P})$ or $\mathbf{S} \in pds_s(X_{j,s}, X_{i,t}, \mathcal{P})$ and $X_{i,t} \perp\!\!\!\perp X_{j,s} | \mathbf{S}$ according to (Test, α)

$$\textbf{then} \begin{cases} \text{Delete edge } X_{i,t} \circ\!\!-\!\!\circ X_{j,s} \text{ from } \mathcal{P}. \\ \text{Delete edge } X_{m,a} \circ\!\!-\!\!\circ X_{n,b}\ \forall (X_{m,a}, X_{n,b}) \in \textit{hom}(X_{i,t}, X_{j,s}, \mathcal{P}). \\ \text{Let } \textit{sepset}(X_{i,t}, X_{j,s}) = \textit{sepset}(X_{i,t}, X_{j,s}) = \mathbf{S}. \end{cases}$$

13. Reorient all edges as $\circ\!\!-\!\!\circ$ and **repeat** steps 10 and 11.
14. Exhaustively apply orientation rules (R1-R10) in [7] to orient remaining endpoints, orienting all homologous pairs similarly.
15. **return** $\mathcal{P}$.

2 Proofs

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

Proof. The proof of local consistency in [3] only requires that the score is decomposible and consistent. That the VECM BIC score is decomposible is true by definition of the score. So, we only need to argue that the VECM BIC score is consistent. Under assumptions A1 and A2, least squares regression is consistent for estimating the VECM parameters, so we have a consistent estimate of $\mathbf{\Sigma}$. The consistency of the VECM BIC score follows from a result in [6]. In that work, the author defines a novel model score called the Posterior Information Criterion (PIC), and shows that the PIC score is consistent for I(1) processes under assumptions weaker than those stated here (p. 768). The PIC score is more complicated than BIC in finite samples, but asymptotically the PIC and BIC are almost equivalent, except that the PIC includes an extra constant factor $\kappa > 0$ multiplying the penalty term, i.e., that PIC is asymptotically $\frac{T}{2} \log |\hat{\mathbf{\Sigma}}| - \kappa \frac{d}{2} \log T$ (p. 776). In fact, $\kappa \geq 1$ for VECM models, since the additional penalty on model dimension comes from double-counting the parameters of the cointegrating matrix $\mathbf{\Pi}$ in a VECM [1, p. 7-8].

The BIC score ranks models in the same order as the PIC score, asymptotically. To see this, consider two models $\mathcal{G}_1$ and $\mathcal{G}_2$. Say PIC ranks $\mathcal{G}_1$ higher than $\mathcal{G}_2$. There are two cases to consider. In the first case, the two models have the same likelihoods. PIC($\mathcal{G}_1$) $>$ PIC($\mathcal{G}_2$) if and only if $0 > (d_1 - d_2)\kappa \log T$, i.e.,

$d_2 > d_1$. The BIC score would then also rank $\mathcal{G}_1$ higher than $\mathcal{G}_2$. In the second case, the models have different likelihoods. $\mathrm{PIC}(\mathcal{G}_1) > \mathrm{PIC}(\mathcal{G}_2)$ if and only if $\frac{T}{2}\log|\hat{\mathbf{\Sigma}}_1| - \frac{T}{2}\log|\hat{\mathbf{\Sigma}}_2| > (d_1 - d_2)\kappa \log T$. Since $\kappa \log T \geq \log T$ for $\kappa \geq 1$, the BIC score would also rank $\mathcal{G}_1$ higher than $\mathcal{G}_2$. Thus, the consistency of the VECM BIC score follows from the consistency of the PIC score. □

For related work on the consistency of the BIC score for order selection (determining p) and cointegration rank selection, see [5] and [2] respectively. We note that future research may explore the prospects of using the PIC score (either in its exact finite sample form, or in asymptotic form) in place of BIC in greedy score-based search.

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 $\mathcal{P}$.*

Proof. Under assumptions A1-3 the VECM BIC score is a decomposable, consistent, and score-equivalent score. The BIC CI test is a consistent test of conditional independence, as a consequence of local consistency of the score. The conclusion follows from the consistency of the SVAR-GFCI algorithm [4]. □

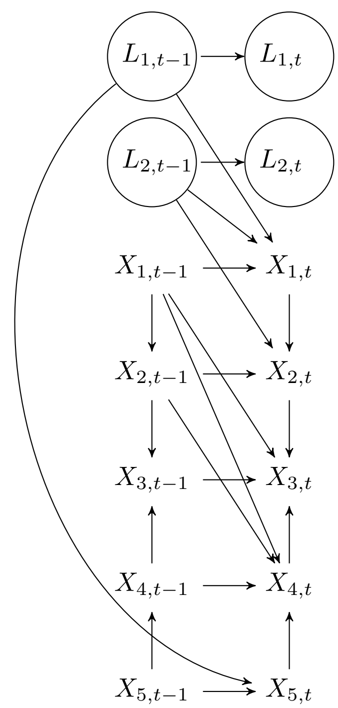

Figure 1: The assumed data-generating process for our nonstationary simulation study. The latent processes are pure random walks.

References

[1] John C Chao and Peter CB Phillips. Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics*, 91(2):227–271, 1999.

[2] Xu Cheng and Peter CB Phillips. Cointegrating rank selection in models with time-varying variance. *Journal of Econometrics*, 169(2):155–165, 2012.

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

[4] 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.

[5] Jostein Paulsen. Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis*, 5(2):115–127, 1984.

[6] Peter CB Phillips. Econometric model determination. *Econometrica: Journal of the Econometric Society*, pages 763–812, 1996.

[7] 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.