Instrumental Processes Using Integrated Covariances

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Instrumental variable methods are often used for parameter estimation in the presence of confounding. They can also be applied in stochastic processes. Instrumental variable analysis exploits moment equations to obtain estimators for causal parameters. We show that in stochastic processes one can find such moment equations using an integrated covariance matrix. This provides new instrumental variable methods, instrumental variable methods in a class of continuous-time processes as well as a unified treatment of discrete- and continuous-time processes.

Keywords: instrumental variables, point processes, linear Hawkes processes, VAR(p), time series, causal inference, recurrent events

1. Introduction

Instrumental variable (IV) techniques have a long history in economics, engineering, and causal inference, even if each field has its own standard formulation of the IV problem (Wright, 1928; Reiersøl, 1941, 1945; Sargan, 1958; Joseph et al., 1961; Wong, 1966; Wong and Polak, 1967). Recent work (Thams et al., 2022) formulates an instrumental variable problem in a (discrete-time) time series model and provides a solution which employs conditional instruments (Brito and Pearl, 2002). Thams et al. (2022) take a *variable-centric* approach in that they identify sets of variables at different lags that satisfy conditions enabling conditional instrumental variable techniques. This paper takes a *process-centric* approach, essentially by integrating out time. The IV methods of this paper therefore only use integrated measures of covariance of stochastic processes. The distinction between variable- and process-centric will be described in more detail in Section 2.

The process-centric approach outlined in this paper is applicable to discrete-time stochastic processes and can also be applied in continuous time as we show using a class of point processes. The estimand is slightly different than in existing methods, however, the estimated parameter is easily interpretable and it gives a simple measure summarizing the strength of the dependence between stochastic processes.

As the paper uses both discrete- and continuous-time models, we only use the term *time series* to refer to stochastic processes in discrete time. The paper is structured as follows. Section 2 describes a classical instrumental variable problem as well as the variable-centric and process-centric approaches to IV estimation in time series. Section 3 describes the classes of stochastic processes that we use in this paper as well as the causal estimands that our IV equations identify. Section 4 describes IV methods in both linear Hawkes processes and vector-autoregressive time series. In both, we use an integrated covariance matrix to obtain new IV results and there is a strong conceptual similarity between the two, even though the interpretation of the parameters depends on the model class. We also generalize the results in the time series setting to allow more general confounding and more general instrumental processes (Section 5). Section 6 discusses estimation.

© 2023 .

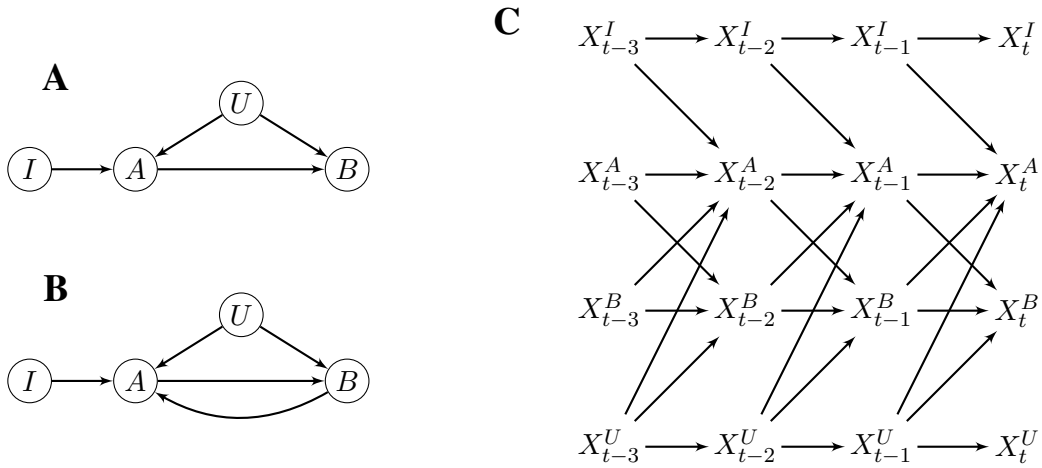

Figure 1: Graphical representations of examples in Section 2. **A**: Graph representing the IV model in Example 1. Each node (I, A, U, B) represents a random variable in the model. **B**: Graph representing the time series IV model in Section 2. Each node represents a coordinate process. **C**: An *unrolled* version of **B** (Danks and Plis, 2013) representing the time series IV model. Each node represents a random variable. The analogous graph with a node for every random variable in the time series is known as the *full time graph* (Peters et al., 2013).

2. Instrumental variable methods

In this section, we give examples of a classical IV problem, that is, using variables that are not indexed by time. We then compare this to a simple vector-autoregressive model of order 1, VAR(1). In this model, we explain the *variable*- and *process*-centric approach to IV estimation and show how the integrated covariance enables IV estimation. We assume zero-mean random variables as the generalization is straightforward.

Example 1 (Classical IV) *Assume we have observable, zero-mean random variables I, A, B and*

$$B = \phi A + \varepsilon$$

where ε is a zero-mean random variable and we wish to estimate $\phi \in \mathbb{R}$. If ε and A are correlated, then least-squares estimation is biased. If I is uncorrelated with ε and $E(AI) \neq 0$, then we say that I is an instrumental variable. *Multiplying by I, and taking expectations, we obtain*

$$E(BI) = \phi E(AI). \tag{1}$$

*This moment equation identifies the parameter ϕ as $E(AI) \neq 0$, even if A and ε are correlated, for instance, due to an unobserved confounder, U, see Figure 1**A**.*

From the above it is clear that the parameter ϕ is in fact identified from the covariance matrix of the vector $(A, B, I)^T$, that is, the observed covariance matrix is sufficient for IV estimation. The

central idea of this paper is to use a different observable matrix in a stochastic process setting which is also sufficient for IV estimation. The next example illustrates this in a simple manner.

Example 2 (Time series IV) *We can instead consider a time series model with a similar structure as in Example 1. Let $X_t = (X_t^I, X_t^A, X_t^B, X_t^U)^T$ such that X_t^U is unobserved and processes $X_t^I, X_t^A, X_t^B, X_t^U$ are all one-dimensional and zero-mean. For simplicity, we assume X_t to be a vector-autoregressive process of order 1, VAR(1),*

$$X_t = \Phi X_{t-1} + \varepsilon_t$$

where ε_t are identically distributed and independent random vectors with independent entries. The matrix Φ has the following structure,

$$\Phi = \begin{bmatrix} \Phi_{II} & 0 & 0 & 0 \\ \Phi_{AI} & \Phi_{AA} & \Phi_{AB} & \Phi_{AU} \\ 0 & \Phi_{BA} & \Phi_{BB} & \Phi_{BU} \\ 0 & 0 & 0 & \Phi_{UU} \end{bmatrix}$$

*We assume that each entry of Φ is nonzero if it is not explicitly zero above. There is a graphical representation of this process in Figure 1**B** where $Z \to Y$ if and only if $\Phi_{YZ} \neq 0$ for $Z, Y \in \{I, A, B, U\}$. Graph **C** is an* unrolled *version (Danks and Plis, 2013) of graph **B** where the nodes represent random variables and $X_{t-1}^Z \to X_t^Y$ if and only if $\Phi_{YZ} \neq 0$.*

If we were to apply the approach from Example 1, we could use X_{t-2}^I as an instrument to identify the parameter Φ_{BA} which corresponds to the edge $X_{t-1}^A \to X_t^B$ and write

$$\begin{aligned} X_t^B &= \Phi_{BA} X_{t-1}^A + \Phi_{BB} X_{t-1}^B + \Phi_{BU} X_{t-1}^U + \varepsilon_t^B \\ &= \Phi_{BA} X_{t-1}^A + \bar{\varepsilon}_t^B \\ E(X_t^B X_{t-2}^I) &= \Phi_{BA} E(X_{t-1}^A X_{t-2}^I) + E(\bar{\varepsilon}_t^B X_{t-2}^I) \end{aligned}$$

where $\bar{\varepsilon}_t^B = \Phi_{BB} X_{t-1}^B + \Phi_{BU} X_{t-1}^U + \varepsilon_t^B$. Thams et al. (2022) (Proposition 6) show that using the moment equation in (1), with $I = X_{t-2}^I, A = X_{t-1}^A, B = X_t^B$, does not lead to consistent estimation of Φ_{BA} when both Φ_{II} and Φ_{BB} are nonzero. Therefore, naive application of classical IV methods will not give consistent estimation in this problem. This can be explained by the fact that there are confounding paths going back in time, e.g., $X_{t-2}^I \leftarrow X_{t-3}^I \to X_{t-2}^A \to X_{t-1}^B \to X_t^B$, corresponding to the fact that $E(\bar{\varepsilon}_t^B X_{t-2}^I)$ is not necessarily zero.

Thams et al. (2022) instead provide consistent estimators of Φ_{BA} using conditional *instrumental variables, using a conditional version of the moment equation in Equation (1). In this case, I_{t-2} is a conditional instrument for the parameter Φ_{BA}* conditionally *on X_{t-3}^I (Thams et al., 2022, Theorem 7).*

The conditional instrumental variable approach is *variable-centric* in the sense that it identifies finite sets of variables that satisfy assumptions of a conditional instrumental variable method as in the above example. In this paper, we take a different approach which will also provide a solution to

the instrumental variable problem in the above example. Instead of looking at covariances of single variables, e.g., between X_t^B and X_{t-2}^I, we use an integrated measure of covariance, summing out temporal dependence. Taking this point of view, we arrive at an unconditional instrumental variable method in the above example, and we say that this is a *process-centric* approach as it uses the integrated covariance. The rest of this section describes this idea in the VAR(1)-example, though we no longer require X_t^I, X_t^A, X_t^B, and X_t^U to be one-dimensional.

Assume that the observed variables are mean-zero and that the largest absolute value of the eigenvalues of Φ is strictly less than one. In the model from Example 2, we see that for a fixed t, and using that ε_{t-j} and ε_{t+i-k} are independent unless $k = l - j$,

$$\begin{aligned} C &= \sum_{k=-\infty}^{\infty} E\left(X_t X_{t+k}^T\right) = \sum_{k=-\infty}^{\infty} E\left(\left(\sum_{j=0}^{\infty} \Phi^j \varepsilon_{t-j}\right)\left(\sum_{l=0}^{\infty} \Phi^k \varepsilon_{t+k-l}\right)^T\right) \\ &= \left(\sum_{j=0}^{\infty} \Phi^j\right) \Theta \left(\sum_{j=0}^{\infty} \Phi^j\right)^T = (I-\Phi)^{-1}\Theta(I-\Phi)^{-T} \end{aligned} \tag{2}$$

where Θ is the diagonal covariance matrix of ε_t. This result also follows from standard VAR-process results (Brockwell and Davis, 1991, see the covariance matrix generating function). We will say that Equation (2) is the *integrated covariance equation*. We will see that the linear Hawkes model and more general time series models also satisfy this equation when the parameter matrices are given the correct interpretations. There is also a clear similarity with the parametrization of the observed covariance of a linear structural equation model as noted by Mogensen (2022) in the linear Hawkes model. Therefore, more general identification results from cyclic linear structural equation models may be used (Mogensen, 2022).

One can straightforwardly show that $(I-\Phi_{BB})^{-1}\Phi_{BA} = C_{BI}(C_{AI})^{-1}$ when C_{AI} is invertible, thus identifying the matrix $(I-\Phi_{BB})^{-1}\Phi_{BA}$ of *normalized parameters* (Subsection 3.3). This matrix has a clear causal interpretation in both settings, see Subsection 3.3, summarizing the direct influence of one subprocess on another. In the following sections, we show that this approach also applies to more general time series models as well as to linear Hawkes processes, a class of multivariate, continuous-time point processes.

3. Probabilistic Models

In this section, we introduce the class of *linear Hawkes process* as well as the time series models that we are using in this paper. We also show that they satisfy a version of Equation (2) which enables the instrumental variable methods of Section 4. Finally, we describe *normalized parameters* in more detail as these will constitute our estimands.

3.1. Linear Hawkes Processes

A *linear Hawkes process* is a certain kind of *point process*. We give a short introduction here, see also, e.g., Laub et al. (2015); Daley and Vere-Jones (2003). We consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ where $(\mathcal{F}_t)$ is a filtration and an index set $V = \{1, 2, \ldots, n\}$. For $i \in V$, there is a sequence of random event times $\{T_k^i\}_{k\in\mathbb{Z}}$ such that $T_k^i < T_{k+1}^i$ almost surely. We define a counting

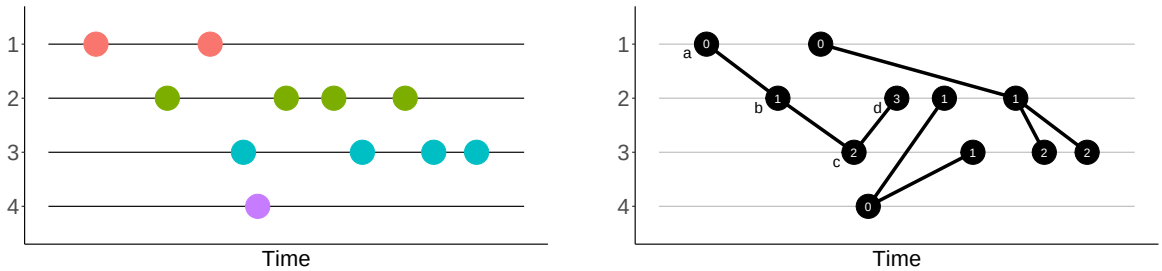

Figure 2: Example data from a four-dimensional linear Hawkes process. Left: Example observed data. Color and vertical placement indicate coordinate process (1, 2, 3, or 4) of the event. Horizonal placement indicates time of the event. Right: The linear Hawkes process can be generated as a *cluster process* where each event may spark future child events, indicated here with line segments. These parent-child relations are unobserved. In the cluster with labelled events (a, b, c, d), event b is in the first generation from event a while event d is in the third generation from event a. We say that b is a *child* of a (direct descendant).

process N_t^i such that $N_t^i - N_s^i = \sum_k \mathbb{1}_{s<T_k^i\leq t}$. Furthermore, we assume that two events cannot occur simultaneously in the multivariate point process. A linear Hawkes process can be defined by imposing constraints on the *conditional intensities*, λ_t^i. These are stochastic processes and satisfy

$$\lambda_t^i = \lim_{h\downarrow 0} \frac{1}{h} P(N_{t+h}^i - N_t^i = 1 \mid \mathcal{F}_t)$$

where $\mathcal{F}_t$ represent the history of the process until time point t. A multivariate *linear Hawkes process* is a point process such that

$$\lambda_t^j = \mu_j + \sum_{i=1}^{n} \int_{-\infty}^{t} \phi_{ji}(t-s) dN_s^i$$

for a positive constant μ_j and nonnegative functions ϕ_{ji} which are zero outside $(0, \infty)$. We define Φ to be the $n \times n$ matrix such that $\Phi_{ji} = \int_{-\infty}^{\infty} \phi_{ji}(s) ds$. See Figure 2 for an illustration of data observed from a linear Hawkes process. When A is a square matrix, we let $\rho(A)$ denote its *spectral radius*, that is, the largest absolute value of its eigenvalues. We assume that $\rho(\Phi) < 1$ in which case we can assume the linear Hawkes process to have stationary increments (Jovanović et al., 2015).

We define the integrated covariance in this setting,

$$C_{ij} dt = \int_{-\infty}^{\infty} E(dN_t^i dN_{t+s}^j) - E(dN_t^i) E(dN_{t+s}^j) ds. \tag{3}$$

We also define $\Lambda_i dt = E(dN_t^i)$ and let Θ denote the diagonal matrix such that $\Theta_{ii} = \Lambda_i$. It holds that $C = (I - \Phi)^{-1} \Theta (I - \Phi)^{-1}$ (Achab et al., 2017). This is the same equation as in the VAR(1)-case in Section 2, even though interpretations of the parameter matrices Φ and Θ differs.

3.1.1. Cluster interpretation

Above we introduced the linear Hawkes process as a point process with conditional intensities of a certain type. It is, however, possible to give an equivalent definition using the so-called *cluster representation* (Jovanović et al., 2015). We will give a very short description here. For each $i \in V$, a set of generation-0 events are generated from a homogeneous Poisson process with rate μ_i. Each of these events create a *Hawkes cluster* which is generated in the following way. From an generation-n event at time s of type i (coordinate process i), generation-$(n+1)$ events of type j are generated from an inhomogenous Poisson process started at s with rate $\phi_{ji}(t-s)$, $t > s$. This construction is repeated. The superposition of all clusters form a linear Hawkes process. Note that only event types and time points are observed while generation and parent-child relations of an event are unknown when observing data from a linear Hawkes process.

The cluster interpretation also provides a straightforward interpretation of the entries of Φ. The entry Φ_{ji} is the expected number of direct j-children from an i-event. In general, $(\Phi^k)_{ji}$ is the expected number of j-events from an i-event in the k'th generation from the i-event. We define $R = (I - \Phi)^{-1} = \sum_{k=0}^{\infty} \Phi^k$. R_{ji} is the total number of j-descendants on a cluster rooted at an i-event. Note that R is well-defined and that the infinite sum converges due to the assumption on the spectral radius of Φ (Jovanović et al., 2015). See Figure 2 for an example of direct/indirect descendant events.

3.2. Time Series

Let $X_t = (X_t^1, \ldots, X_t^n)^T$ be a multivariate time series in discrete time, $t \in \mathbb{Z}$. We assume that X_t is stationary and that $E((X_t^i)^2) < \infty$ for all i and t. We say that X_t is a VAR(p)-process if

$$X_t = \sum_{i=1}^{p} \Phi_i X_{t-i} + \varepsilon_t \tag{4}$$

where the ε-process is mean-zero and stationary, ε_t and ε_s are uncorrelated for $s \neq t$, and $E(\varepsilon_t \varepsilon_t^T) = \Theta$. Define $\Phi(z) = I - \Phi_1 z - \ldots - \Phi_p z^p$. We assume that $\det(\Phi(z)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. This means that there exists a unique stationary solution to the VAR(p)-equation (Brockwell and Davis, 1991, Theorem 11.3.1) and we assume throughout that we observe a stationary time series. We use the notation $\Phi = \sum_{i=1}^{p} \Phi_i$. The above assumption on $\Phi(z)$ implies that $I - \Phi$ is invertible. The entries of the matrix $(I - \Phi)^{-1}$ are sometimes called *long-run effects* (Lütkepohl, 2005). We also assume that $I - \Phi_{BB}$ is invertible when needed, see also Subsection 3.3.

We define again the *integrated covariance* in this model class

$$C = \sum_{i=-\infty}^{\infty} E(X_t X_{t+i}^T) - E(X_t) E(X_{t+i})^T.$$

C is well-defined since the sum converges (Brockwell and Davis, 1991, p. 420). Brockwell and Davis (1991) (Section 11.2) discuss estimation of the terms $E\left(X_t X_{t+i}^T\right)$. The matrix C is independent of t due to stationarity. One should also note that the matrix C equals 2π times the spectral density matrix of $X_t - E(X_t)$ at 0.

We saw in Section 2 that the integrated covariance equation holds for VAR(1)-processes and we can extend this result to VAR(p)-processes. This is done in Appendix A. We obtain the same

integrated covariance equation, however, Φ is now the sum of the direct effects for each lag $k = 1, \ldots, p$, that is, $\Phi = \sum_{k=1}^{p} \Phi_k$. Again, this result is also implied by textbook results on time series (Brockwell and Davis, 1991, p. 420).

We note that Φ in the VAR(p)-case may have negative entries which is different from the linear Hawkes case. This means that some results that apply in the linear Hawkes setting do not hold in VAR(p)-time series, e.g., in relation to marginalization (Mogensen, 2022; Hyttinen et al., 2012).

3.3. Normalized parameters

The entries of the parameter matrix Φ have an intuitive interpretation in both model classes. However, in general we will not be able to identify these parameters with the methods in this paper, see Example 10. Instead, we will aim to identify the entries of the *normalized* parameter matrix. These parameters also have a useful interpretation as we will explain.

We use I_n to denote the identity matrix of dimension n and I_b to denote the identity matrix of dimension $|B|$ for a finite set B.

Definition 3 (Normalized parameters) *Consider a pair of matrices (Φ, Θ) that solve the integrated covariance equation. We say that they are* normalized *if $\Phi_{ii} = 0$ for all i.*

Say we consider any pair (Φ, Θ) and wish to normalize it. We define D to be the diagonal matrix such that $D_{ii} = (1 - \Phi_{ii})^{-1}$. Note that $\Phi_{ii} \neq 1$ due to the assumptions on Φ (Subsections 3.1 and 3.2). Then D is invertible and

$$\begin{aligned} C &= (I_n - \Phi)^{-1}\Theta(I_n - \Phi)^{-T} = (D(I_n - \Phi))^{-1}D\Theta D(D(I_n - \Phi))^{-T} \\ &= (I_n - \bar{\Phi})^{-1}\bar{\Theta}(I_n - \bar{\Phi})^{-T} \end{aligned}$$

We see that $(\bar{\Phi})_{ji} = \Phi_{ji}/(1 - \Phi_{jj})$ and that $\bar{\Phi}$ has zeros on the diagonal and therefore $(\bar{\Phi}, \bar{\Theta})$ is normalized. If $\rho(\Phi) < 1$ and the entries of Φ are nonnegative, then this will also be the case for $\bar{\Phi}$ (Mogensen, 2022). This means that the normalized parameters are also within the Hawkes parameter space.

The interpretation of the normalized parameters depend on the model class. In the linear Hawkes process, $\bar{\Phi}_{ji}$ is the expected number of j events on a cluster rooted at an i event counting only subtrees of the form $i - j - j - \ldots - j$ for any number of j events. This is thus the expected number of direct j events from an injected i event when also counting subsequent 'self-events' $j - j$. In the time series case, we see that $(\Phi_{jj})^k\Phi_{ji}$ is the partial causal effect corresponding to the path $X_t^i \to X_{t+1}^j \to X_{t+2}^j \to \ldots \to X_{t+k+1}^j$. If $|\Phi_{jj}| < 1$, we have $\Phi_{ji}/(1 - \Phi_{jj}) = \sum_{k=0}^{\infty}(\Phi_{jj})^k\Phi_{ji}$ and the normalized parameter is therefore the sum of the partial effects (Tian, 2004) along all paths of the type $X_t^i \to X_{t+1}^j \to X_{t+2}^j \to \ldots \to X_{t+k+1}^j$ and a measure of the causal influence of the variable X_t^i on the entire future of process j, counting the direct effect as well as subsequent self-effects. In both cases, the normalized parameters are seen to represent an easily interpretable causal quantity.

We will also use quantities of the type $(I_b - \Phi_{BB})^{-1}\Phi_{BA}$ which is a multivariate version of the above. The interpretation generalizes in a straightforward manner to this case. We see that $\Phi_{BB}^i\Phi_{BA}$ are the partial effects (Tian, 2004) from X_t^A to X_{t+i+1}^B corresponding to paths $A \to B \to B \to \ldots \to B$. If $\rho(\Phi_{BB}) < 1$, this means that $(I - \Phi_{BB})^{-1}\Phi_{BA} = \sum_{k=0}^{\infty}\Phi_{BB}^k\Phi_{BA}$ is an aggregate

causal effect from X_t^A to $\{X_{t+j}^B\}_{j\geq 1}$ taking only paths of the type $A \to B \to B \to \ldots \to B$ into account. In this sense, it is a direct effect of A at time t on the entire future B-process counting the direct effect $X_t^A \to X_{t+1}^B$ and subsequent self-effects within B. Therefore, this is a natural quantification of the effect of subprocess A on subprocess B when taking a stochastic process point of view. The condition that $\rho(\Phi_{BB}) < 1$ is a natural requirement for the above interpretation of normalized parameters as this means that the marginal B-time series is 'stable' in itself. This ensures that the induced self-effects are finite.

Example 10 in Appendix C shows that from a normalized pair, (Φ, Θ), every diagonal matrix, D_{ii}, such that $D_{ii} \neq 1$, provides us with a different pair $(\bar{\Phi}, \bar{\Theta})$ solving the same integrated covariance equation as the original pair. If $\rho(\Phi) < 1$ and the entries are nonnegative and we let $0 < D_{ii} < 1$, then the same holds for $\bar{\Phi}$. This means that in both the time series case and the linear Hawkes case we may find infinitely many pairs $(\bar{\Phi}, \bar{\Theta})$ that solve the equation. This needs a short argument in the linear Hawkes case to ensure that $\rho(\bar{\Phi}) < 1$. In the time series case, we need to argue that $I - \bar{\Phi}_{BB}$ is also invertible. These arguments are provided in Example 10 in Appendix C. Hyttinen et al. (2012) provide similar arguments in the context of cyclic linear structural equation models.

3.4. Graphical representation

One may use graphs to represent assumptions that are sufficient for IV analysis. These graphs are defined for linear Hawkes models and VAR(p)-models below.

Definition 4 (Causal graph) *Let $\mathcal{G}$ be a directed graph on nodes V and with edge set E. In the linear Hawkes case, we say that $\mathcal{G}$ is the* causal graph *of the process if $i \to j$ is in E, $i \neq j$, if and only if $\Phi_{ji} \neq 0$. In the VAR(p) model, we say that $\mathcal{G}$ is the* causal graph *of the process if $i \to j$ is in E, $i \neq j$, if and only if there exists k such that $(\Phi_k)_{ji} \neq 0$.*

Note that the causal graph does not contain loops, that is, edges $i \to i$. When identifying *normalized* parameters, loops are inconsequential as the normalization removes self-effects and adjusts the other parameters to retain the integrated covariance.

We say that a process is *exogenous* if it has no parents in the causal graph. We say that a subset of processes, $I \subseteq V$, are *exogenous* if there are no $\alpha \notin I$ and $\beta \in I$ such that $\alpha \to \beta$ in the causal graph. Note that there could be edges between processes in an exogeneous set, I, only not from processes $V \setminus I$ and into I.

4. Instrumental Processes

The exact statements and the proofs of the IV results are different between the two model classes when we allow for a more general confounding in the time series case. For this reason, we first describe the results for a linear Hawkes process and a VAR(p)-process as these are completely analogous. However, the interpretation of the parameters differ as seen in Section 3. We then describe how to generalize this in the time series case. In this section, *process* refers to either a VAR(p)-process or a linear Hawkes process.

This section uses the algebraic equation in (2) to define instrumental processes that allow us to identify *normalized* causal parameters (see Definition 3). Mogensen (2022) notes that the parametrization of the integrated covariance is similar to the parametrization of the covariance of a

linear structural equation model for which there are several identification results, see, e.g., Foygel et al. (2012); Chen (2016); Weihs et al. (2018). We will not use this connection directly and therefore we refer to that paper for a detailed explanation. One should note that identification results from linear SEMs could be used to obtain some of the results of this paper. However, we take a more direct approach which is closer to other IV work. Furthermore, this approach also makes the needed assumptions explicit whereas identification results are often only generic, that is, hold outside a measure-zero set of parameters.

Even though the results in this section are similar in spirit to other IV work, we use the matrix C directly, and not a set of random variables. C is easily seen to be similar to a covariance matrix, but it is an aggregate measure of covariance between processes rather than the covariance of a set of observed random variables.

We give first a univariate definition of an instrumental process which leads to an identification result. Then we define a multivariate instrumental process and state the corresponding identification result. The univariate definition and result are naturally implied by the multivariate result. However, we include them in order to present the simplest possible setting first.

We are now ready to define what we mean by an instrumental process. The symbol ι will throughout the paper denote an instrumental process (instrumental for the effect from α to β). The symbol I will denote an instrumental set (multiple instruments, instrumental for the effect from the set A to the set B), that is, $I, A, B \subseteq O \subseteq V$ where O is the set of observed processes. We assume that I, A, and B are disjoint.

Definition 5 *We say that ι is an* instrumental process *for $\alpha \to \beta$ in the causal graph $\mathcal{G}$ if ι is exogenous, every directed path from ι to β includes the edge $\alpha \to \beta$, and $C_{\alpha\iota} \neq 0$.*

Theorem 6 *Let $\iota, \alpha, \beta \in O$ and $\mathcal{G} = (V, E)$ is the causal graph, $V = O \dot{\cup} U$. If ι is an instrumental process for $\alpha \to \beta$, then $(1 - \Phi_{\beta\beta})^{-1}\Phi_{\beta\alpha}$ is identified by $C_{\beta\iota}/C_{\alpha\iota}$.*

Example 7 (Instrumental process) *In this example, we show that the classical IV graph also allows an IV analysis in this setting. Say we have a four-dimensional linear Hawkes process such that the causal graph is as shown in Figure 3**A** and process 4 is unobserved. Then 1 is an instrument for the normalized effect from 2 to 3. Theorem 6 gives that*

$$C_{3,1}/C_{2,1}$$

identifies this effect.

4.1. Multiple instruments

As in other instrumental variable frameworks, we may consider using *multiple instruments* in case there are multiple processes that are instrumental for the same (collection of) effects.

Definition 8 *Let $I, A, B \subseteq O$. We say that a set of processes, I, is an* instrumental set *for $A \to B$ in the causal graph $\mathcal{G}$ if I is exogenous, every directed path from I to B includes an edge in $A \to B$, and C_{AI} has full row rank.*

Theorem 9 (Multiple instruments (just identified)) *Let $I, A, B \subseteq O$ be disjoint and non-empty sets such that I is an instrumental process for the effect from A and B and assume that $|A| = |I|$. In this case, $(I_b - \Phi_{BB})^{-1}\Phi_{BA}$ is identified by $C_{BI}(C_{AI})^{-1}$.*

Figure 4 gives an example of a graphical structure with a multivariate instrumental process.

Figure 3: **A**: Instrumental process example. Process 4 is unobserved (indicated by the square). Process 1 (ι) may serve as an instrumental process to estimate the normalized effect from 2 (α) to 3 (β). **B**: The graph in **A** is only one possible explanation. More generally, in any graph for which **B** is the *latent projection* (Verma and Pearl, 1990; Richardson et al., 2017) of the causal graph the same instrumental process technique as used in **A** would work.

4.2. Overidentification

Consider instead the case where $|A| < |I|$, that is, overidentification. In this case, C_{AI} is not invertible. Let C_{AI}^{-} be a right inverse, that is, C_{AI}^{-} is an $|I| \times |A|$ matrix such that $C_{AI}C_{AI}^{-} = I_a$. Such a matrix exists as C_{AI} has full row rank by assumption. Note that from this assumption it also follows that R_{AI} has full row rank as $\mathrm{rank}(AB) \leq \mathrm{rank}(A)$ for matrices A and B. We see that $R_{AI}^{-} = \Lambda_{II}R_{II}C_{AI}^{-}$ is a right-inverse of R_{AI}. Then

$$R_{BI}R_{AI}^{-} = C_{BI}C_{AI}^{-}$$

The rest of the proof of Theorem 9 holds also in this case, showing that any choice of right-inverse of C_{AI} leads to identification of the normalized parameters. Note that choosing a specific right-inverse of C_{AI} specifies a choice of right-inverse of R_{AI} as well – this specific right-inverse is then used throughout the proof.

When W is a positive definite weight matrix then $C_{AI}WC_{AI}^{T}$ is invertible using the fact that C_{AI} has full rank. We see that the matrix $WC_{AI}^{T}(C_{AI}WC_{AI}^{T})^{-1}$ is a right-inverse of C_{AI}. This motivates using

$$C_{BI}WC_{AI}^{T}(C_{AI}WC_{AI}^{T})^{-1}$$

as an estimate in the overidentified setting by plugging in estimated entries of C. See also Thams et al. (2022) and Hall (2005).

5. Time series, general case

We now argue that the above methods apply to time series models under far more general assumptions. In this section, we dispense with the graphical conditions and assume that X^B is given as below. The same procedure works if there are several lagged X^A- and X^B-vectors in the expression below (Appendix B). One should also note that we are not making any assumptions about the marginal distribution of X^I except as required for the existence of the relevant entries of C and the full rank of C_{AI}.

We assume

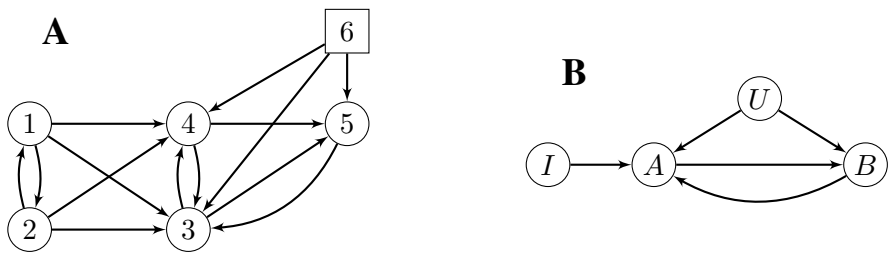

Figure 4: **A**: Multivariate instrumental process example. Process 6 is unobserved (indicated by the square). Processes 1 and 2 (I) may serve as an instrumental set to estimate the normalized effect from 3 and 4 (A) to 5 (B). **B**: This graph is a simplified version of **A**. We collapse processes 1 and 2 into a single node and processes 3 and 4 into another node, defining sets $I = \{1, 2\}, A = \{3, 4\}, B = \{5\}, U = \{6\}$ where U is unobserved. For $X, Y \in \{I, A, B, H\}$, we include edges $X \to Y$ if and only if $x \to y$ for some $x \in X$ and $y \in Y$. This recovers the 'univariate' IV structure from Figure 1**B**. Thams et al. (2022) use this graphical representation as well as the full time graphs as described below Figure 1.

$$X_t^B = \Phi_{BA} X_{t-1}^A + \Phi_{BB} X_{t-1}^B + g_B(\ldots, X_{t-2}^U, X_{t-1}^U, \varepsilon_t^B)$$

such that $\varepsilon_t = (\varepsilon_t^I, \varepsilon_t^A, \varepsilon_t^B)$ are independent random variables and also independent of X^U. These assumptions correspond to more flexible confounding than above. We also assume that X_t^I is independent of U and ε^B for all t and that $(I - \Phi_{BB})$ is invertible. Assume that X_t^I, X_t^A, X_t^B are mean-zero. In this section, we will simply assume that the relevant entries of C are well-defined, though one can use conditions such as those in the VAR(p)-models to ensure this, noting that we only need the existence of C_{BI} and C_{AI}. We may write

$$\begin{aligned} E\Big(X_{t+i}^B (X_t^I)^T\Big) &= E\Big((\Phi_{BA} X_{t+i-1}^A + \Phi_{BB} X_{t+i-1}^B + g_B(\ldots, X_{t+i-2}^U, X_{t+i-1}^U, \varepsilon_t^B))(X_t^I)^T\Big) \\ &= \Phi_{BA} E\Big(X_{t+i-1}^A (X_t^I)^T\Big) + \Phi_{BB} E\Big(X_{t+i-1}^B (X_t^I)^T\Big). \end{aligned}$$

We sum over i in the above expression,

$$\begin{aligned} \sum_{k=-\infty}^{\infty} E\Big(X_{t+k}^B (X_t^I)^T\Big) &= \sum_{k=-\infty}^{\infty} \Phi_{BA} E\Big(X_{t+k-1}^A (X_t^I)^T\Big) + \sum_{k=-\infty}^{\infty} \Phi_{BB} E\Big(X_{t+k-1}^B (X_t^I)^T\Big) \\ C_{BI} &= \Phi_{BA} C_{AI} + \Phi_{BB} C_{BI} \end{aligned}$$

and isolating C_{BI} we obtain $C_{BI} = (I - \Phi_{BB})^{-1} \Phi_{BA} C_{AI}$. If C_{AI} is has full row rank this gives identification of the matrix $(I - \Phi_{BB})^{-1} \Phi_{BA}$.

6. Estimation

In order to use the instrumental process framework above, one can estimate the relevant entries of the integrated covariance matrix and then plug in the estimated covariances to obtain parameter estimates. Achab et al. (2017) describe how to estimate cumulants of linear Hawkes process. We sketch their approach below. We assume that we observe a linear Hawkes process over the interval $[0, T]$ and that there exists $H > 0$ such that restricting the integration in Equation (3) to $[-H, H]$ introduces only a negligible error. As pointed out by Achab et al. (2017), this is reasonable if the support of ϕ_{ji} is small compared to H and the spectral radius of Φ is sufficiently small. Given a realization of a stationary linear Hawkes process on $[0, T]$ let $p_i = \{t_1^i, \ldots, t_{m_i}^i\} \subset [0, T]$ be the observed event times of process i. The following are estimators of the first- and second-order cumulants,

$$\hat{\Lambda}_i = \frac{1}{T} \sum_{\tau \in p_i} 1, \qquad \hat{C}_{ij} = \frac{1}{T} \sum_{k=1}^{m_i} \left(N^j_{t_k^i + H} - N^j_{t_k^i - H} - 2H\hat{\Lambda}^j \right)$$

In the above, N_t^i refers to the observed counting process corresponding to process i, that is, $N_t^i = 0$ for $t < 0$ and in general $N_t^i = \sum_{k=1}^{m_i} \mathbb{1}_{t_k^i \leq t}$. As noted by Achab et al. (2017), there is a bias in the estimation of the integrated covariance, however, it is found to be negligible. Achab et al. (2017) (Theorem 2.1 and Remark 1) show asymptotic consistency assuming that $H_T \to \infty$ and $H_T^2/T \to 0$ where H_T is the value of the parameter H used when observing the process on the interval $[0, T]$.

To estimate the matrix C in the time series case, one may use the relation to the spectral density matrix of the time series, see, e.g., Hansen (1982); Brillinger (1981). One may also use the connection to *long-run covariance* to obtain consistent estimation of C, see, e.g., Andrews (1991); Andrews and Monahan (1992).

7. Conclusion

The instrumental variable methods in this paper provide a moment equation for time series models which avoids using a conditional moment equation as in Thams et al. (2022). They also make minimal assumptions on the marginal distribution of the instrumental process. On the other hand, they involve an integral or an infinite sum which needs to be estimated when applying the method. One should also note that our estimands are slightly different than those of Thams et al. (2022). As shown the estimands in this paper do have a simple causal interpretation, however.

The integrated covariance approach also allows a unified treatment of IV methods in time series (discrete-time) and continuous-time processes as illustrated by the application to the continuous-time linear Hawkes processes. It is clearly of interest to extend this framework to more general classes of continuous-time processes. Finally, one should also note that the parametrization of the integrated covariance can be used to obtain other identification results than the instrumental variable results in this paper.

References

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Donald WK Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

Donald WK Andrews and J Christopher Monahan. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica: Journal of the Econometric Society*, pages 953–966, 1992.

David R Brillinger. *Time Series: Data Analysis and Theory*. Holden Day, Inc., San Francisco, 1981.

Carlos Brito and Judea Pearl. Generalized instrumental variables. In *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer, New York, 2nd edition, 1991.

Bryant Chen. Identification and overidentification of linear structural equation models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, 2016.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, New York, 2nd edition, 2003.

David Danks and Sergey Plis. Learning causal structure from undersampled time series. In *NIPS 2013 Workshop on Causality*, 2013.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, 40(3):1682–1713, 2012.

Alastair R Hall. *Generalized method of moments*. Oxford University Press, 2005.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

P Joseph, J Lewis, and J Tou. Plant identification in the presence of disturbances and application to digital adaptive systems. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry*, 80(1):18–24, 1961.

Stojan Jovanović, John Hertz, and Stefan Rotter. Cumulants of Hawkes point processes. *Physical Review E*, 91(4):042802, 2015.

Patrick J Laub, Thomas Taimre, and Philip K Pollett. Hawkes processes. *arXiv:1507.02822*, 2015.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, 2005.

Søren Wengel Mogensen. Equality constraints in linear Hawkes processes. In *Proceedings of the 1st Conference on Causal Learning and Reasoning (CLeaR)*, 2022.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, 2013.

Olav Reiersøl. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9(1):1–24, 1941.

Olav Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Stockholms Högskola, 1945.

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv:1701.06686*, 2017.

John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958.

Nikolaj Thams, Rikke Søndergaard, Sebastian Weichwald, and Jonas Peters. Identifying causal effects using instrumental time series: Nuisance IV and correcting for the past. *arXiv:2203.06056*, 2022.

Jin Tian. Identifying linear causal effects. In *Proceedings of the 19th AAAI Conference on Artificial Intelligence*, 2004.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, University of California, Los Angeles, 1990.

Luca Weihs, Bill Robinson, Emilie Dufresne, Jennifer Kenkel, Kaie Kubjas, Reginald McGee II, Nhan Nguyen, Elina Robeva, and Mathias Drton. Determinantal generalizations of instrumental variables. *Journal of Causal Inference*, 6(1), 2018.

K Y Wong. *Estimation of parameters of linear systems using the instrumental variable method*. PhD thesis, University of California, Berkeley, 1966.

Kwan Wong and Elijah Polak. Identification of linear discrete time systems using the instrumental variable method. *IEEE Transactions on Automatic Control*, 12(6):707–718, 1967.

Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Appendix A. Integrated covariance, VAR(p)

First, we rewrite the VAR(p)-process as a VAR(1)-process, Y, with $n \times p$ coordinate processes,

$$Y_t = \Phi_Y Y_{t-1} + \varepsilon_t^Y = \begin{bmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix} Y_{t-1} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The VAR(1)-computations from above still hold which means that the integrated covariance of Y can be written as

$$C_Y = (I - \Phi_Y)^{-1} \Theta_Y (I - \Phi_Y)^{-T}.$$

Note that if 1 is an eigenvalue of Φ_Y, then it corresponds to an eigenvector which is the concatenation of p copies of a single n-vector. This n-vector is also an eigenvector of Φ with eigenvalue 1 which is a contradiction. This means that $I - \Phi_Y$ is invertible. We can use Schur complements and the structure of $(I - \Phi_Y)$ to see that $((I - \Phi_Y)^{-1})_{1:n,1:n} = (I - \Phi)^{-1}$ where $\Phi = \sum_{i=1}^p \Phi_i$. From the sparsity of Θ_Y it follows that

$$C = (C_Y)_{1:n,1:n} = (I - \Phi)^{-1} \Theta (I - \Phi)^{-T}.$$

Appendix B. VAR(p) and general confounding

The reasoning from Section 5 translates to this more complex model with only small adjustments and assuming existence of the relevant entries of C.

$$E(X_{t+i}^B (X_t^I)^T) =$$
$$E\left(\left(\sum_j (\Phi_j)_{BA} X_{t+i-j}^A + \sum_j (\Phi_j)_{BB} X_{t+i-j}^B + g_B(\dots, X_{t+i-2}^U, X_{t+i-1}^U, \varepsilon_{t+i}^U)\right)(X_t^I)^T\right)$$

We sum over i,

$$\begin{aligned} \sum_{i=-\infty}^{\infty} E(X_{t+i}^B (X_t^I)^T) &= \sum_j (\Phi_j)_{BA} \sum_{i=-\infty}^{\infty} E(X_{t+i-j}^A (X_t^I)^T) \\ &+ \sum_j (\Phi_j)_{BB} \sum_{i=-\infty}^{\infty} E(X_{t+i-j}^B (X_t^I)^T) \\ &= \sum_j (\Phi_j)_{BA} \sum_{i=-\infty}^{\infty} E(X_{t+i}^A (X_t^I)^T) \\ &+ \sum_j (\Phi_j)_{BB} \sum_{i=-\infty}^{\infty} E(X_{t+i}^B (X_t^I)^T). \end{aligned}$$

From this it follows that $C_{BI} = (I - \Phi_{BB})^{-1} \Phi_{BA} C_{AI}$.

Appendix C. Normalization

Example 10 *Consider a representation such that Φ is normalized (i.e., has zeros on the diagonal)*

$$C = (I - \Phi)^{-1}\Theta(I - \Phi)^{-T}.$$

For any diagonal matrix such that $D_{ii} \neq 1$ for all i,

$$C = (D(I - \Phi))^{-1}D\Theta D(D(I - \Phi))^{-T} = (I - \bar{\Phi})^{-1}\bar{\Theta}((I - \bar{\Phi}))^{-T}.$$

If $0 < D_{ii} < 1$ and $\rho(\Phi) < 1$, we have that $\rho(\bar{\Phi}) < 1$. To see this note that $\bar{\Phi} = I - D + D\Phi$. This is a nonnegative matrix and let $\lambda = \rho(\bar{\Phi})$. Then a nonnegative (entrywise) x (and nonzero) can be chosen such that $\bar{\Phi}x = \lambda x$ (Horn and Johnson, 1985, Theorem 8.3.1). Φ and x have nonnegative entries and x is nonzero therefore $\Phi x \geq x$ (the inequalities are to be read entrywise) implies that $\rho(\Phi) \geq 1$ (Horn and Johnson, 1985, Theorem 8.3.2) so $(\Phi x)_i < x_i$ for some i. We have $\lambda x = (I - D + D\Phi)x$ and therefore $(\lambda x)_i < x_i$ so $\lambda < 1$. We see that $D\Theta D$ is positive definite. This shows that we cannot identify unnormalized direct effects from the integrated covariance matrix as every nonzero entry of $\bar{\Phi}$ is different from the corresponding entry of Φ (note the diagonal of $\bar{\Phi}$ is nonzero),

$$\bar{\Phi}_{ii} = 1 - D_{ii} + \sum_k D_{ik}\Phi_{ki} = 1 - D_{ii} > 0$$

and for $i \neq j$,

$$\bar{\Phi}_{ij} = \sum_k D_{ik}\Phi_{kj} = D_{ii}\Phi_{ij} < \Phi_{ij}$$

when $\Phi_{ij} \neq 0$.

For the time series case, note also that

$$\bar{\Phi}_{BB} = I_b - D_{BB} + (D\Phi) = I_b - D_{BB} + D_{BB}\Phi_{BB}$$

and when x is a vector such that $x = \bar{\Phi}_{BB}x$ then

$$x = \bar{\Phi}_{BB}x = (I_b - D_{BB} + D_{BB}\Phi_{BB})x$$

This implies $D_{BB}x = D_{BB}\Phi_{BB}x$ and $x = \Phi_{BB}x$ so 1 is an eigenvalue of Φ_{BB} and therefore $I_b - \Phi_{BB}$ is not invertible which is a contradiction. Therefore 1 is also not an eigenvalue of $\bar{\Phi}_{BB}$ and it follows that $I_b - \Phi_{BB}$ is invertible.

Appendix D. Proofs

Proof [Theorem 6] As ι is exogenous, we have that $C_{\alpha\iota} = R_{\iota\iota}\Theta_{\iota\iota}R_{\alpha\iota}$ and $C_{\beta\iota} = R_{\iota\iota}\Theta_{\iota\iota}R_{\beta\iota}$. From the definition of an instrumental process, we have $C_{\alpha\iota} \neq 0$ and there $R_{\alpha\iota} \neq 0$. Therefore $R_{\beta\iota}/R_{\alpha\iota}$ is identified. We have $R = (I_n - \Phi)^{-1}$ and from the sparsity of R and the fact that $I_n = (I_n - \Phi)R$ it follows that $R_{\beta\iota} = \Phi_{\beta\alpha}R_{\alpha\iota} + \Phi_{\beta\beta}R_{\beta\iota}$. Therefore $R_{\beta\iota}/R_{\alpha\iota} = \Phi_{\beta\alpha}/(1 - \Phi_{\beta\beta})$. ■

Proof [Theorem 9] From exogeneity of I, it holds that $C_{BI} = R_{BI}\Lambda_{II}R_{II}$ and $C_{AI} = R_{AI}\Lambda_{I,I}R_{II}$. In the Hawkes case, let μ be the n-vector such that the i'th entry equals μ_i (the constant from the conditional intensity). Λ is a diagonal matrix such the diagonal equals $R\mu = (\sum_{k=0}^{\infty} G^k)\mu$ (Achab et al., 2017) and therefore Λ_{II} is invertible. In the time series case, it is invertible by assumption. R_{II} is invertible as I is exogeneous and $R_{II} = (I_i - G_{II})^{-1}$. If C_{AI} is invertible, so is R_{AI} (note that they are square as $|A| = |I|$). In that case,

$$R_{BI}(R_{AI})^{-1} = C_{BI}(C_{AI})^{-1}$$

and therefore $R_{BI}(R_{AI})^{-1}$ is identified. From the definition of R, we see that $I_n = (I_n - \Phi)R$ and therefore $R = I_n + \Phi R$. This means that

$$R_{BI} = \sum_{A} \Phi_{BA}R_{AI} = \Phi_{BA}R_{AI} + \Phi_{BB}R_{BI}.$$

The last equality comes from the sparsity of R. We obtain

$$R_{BI}(R_{AI})^{-1} = (I_b - \Phi_{BB})^{-1}\Phi_{BA}.$$

Note that R_{AI} is invertible as noted as above. In the linear Hawkes case, it holds that $\rho(\Phi_{BB}) \leq \rho(\Phi) < 1$ (Horn and Johnson, 1985, Corollary 8.1.20) so $I - \Phi_{BB}$ is also invertible. ■