Sampling distribution for single-regression Granger causality estimators

A. J. Gutknecht*†‡ and L. Barnett‡§

November 22, 2019

Abstract

We show for the first time that, under the null hypothesis of vanishing Granger causality, the single-regression Granger-Geweke estimator converges to a generalised χ^2 distribution, which may be well approximated by a Γ distribution. We show that this holds too for Geweke's spectral causality averaged over a given frequency band, and derive explicit expressions for the generalised χ^2 and Γ-approximation parameters in both cases. We present an asymptotically valid Neyman-Pearson test based on the single-regression estimators, and discuss in detail how it may be usefully employed in realistic scenarios where autoregressive model order is unknown or infinite. We outline how our analysis may be extended to the conditional case, point-frequency spectral Granger causality, state-space Granger causality, and the Granger causality F-test statistic. Finally, we discuss approaches to approximating the distribution of the single-regression estimator under the alternative hypothesis.

1 Introduction

Since its inception in the 1960s, Wiener-Granger causality (GC) has found many applications in a range of disciplines, from econometrics, neuroscience, climatology, ecology, and beyond. In the early 1980s Geweke introduced the standard vector-autoregressive (VAR) formalism, and the Granger-Geweke population log-likelihood-ratio (LR) statistic (Geweke, 1982, 1984). As well as furnishing a likelihood-ratio test for statistical significance, the statistic has been shown to have an intuitive information-theoretic interpretation as a quantitative measure of "information transfer" between stochastic processes (Barnett et al., 2009; Barnett and Bossomaier, 2013). In finite sample, the LR estimator requires separate estimates for the full and reduced VAR models, and as such admits the classical large-sample theory, and asymptotic χ^2 distribution (Neyman and Pearson, 1933; Wilks, 1938; Wald, 1943). However, it has become increasingly clear that the "dual-regression" LR estimator is problematic: specifically, model order selection involves a bias-variance trade-off which may skew statistical inference, impact efficiency, and produce spurious results, including negative GC values (Ding et al., 2006; Chen et al., 2006; Stokes and Purdon, 2017).

An alternative *single-regression* (SR) estimator which obviates the problem has been developed in various forms over the past decade (Dhamala et al., 2008a,b; Barnett and Seth, 2014, 2015; Solo, 2016); but, since the large-sample theory no longer obtains, its sampling distribution has remained unknown until now. In addition to the improved efficiency and reduced bias of the SR estimator (Barnett and Seth, 2014, 2015), knowledge of its sampling distribution under the null hypothesis of vanishing causality would allow to construct novel and potentially superior hypothesis tests, especially in the frequency domain where little is known about the sampling distribution of Geweke's spectral GC statistic[1] (Geweke, 1982, 1984). Closing this gap is thus the central object of the present study.

*Institute of Neuroscience and Medicine (INM-6), Forschungszentrum Juelich, Germany.
†MEG Unit, Brain Imaging Center, Goethe University, Frankfurt am Main, Germany.
‡Sackler Centre for Consciousness Science and Dept. of Informatics, University of Sussex, Falmer, Brighton, UK
§Corresponding author: l.c.barnett@sussex.ac.uk

[1]This applies even in the unconditional case where the Geweke spectral statistic (Geweke, 1982) requires only a single estimate of the full model.

We begin in Section 2 with an overview of the theoretical and technical prerequisites of maximum-likelihood (ML) estimation, large-sample theory, the generalised χ^2 distribution, VAR modelling, and Granger causality estimation. Since SR estimators (both time-domain and frequency-band-limited) are continuous functions of the ML estimator for the full VAR model, the problem of determining their asymptotic distributions can be approached via (a second-order version of) the well-known Delta Method (Lehmann and Romano, 2005), in conjunction with a state-space spectral factorisation technique (Wilson, 1972; Hannan and Deistler, 2012) which allows explicit derivation of reduced-model parameters in terms of the full-model parameters (Barnett and Seth, 2015; Solo, 2016). In Section 3 we show in this way that the asymptotic distributions are generalised χ^2, and derive explicit expressions for the distributional parameters in terms of the parameters of the underlying VAR model.

Unlike under the classical theory, the asymptotic null distributions of the SR estimators have a dependence on the true null parameters, which presents a challenge for statistical inference. These issues are addressed in Section 4, in which we present an asymptotically valid Neyman-Pearson test that accommodates this dependence. We discuss the properties of this test in terms of type I and type II error probabilities. In Section 5 we complete our analysis by addressing the issues of model order selection and potentially infinite model orders. We elaborate on how our analysis might be extended in the future to the conditional case (Geweke, 1984), point-frequency spectral GC (Geweke, 1982, 1984), state-space GC (Barnett and Seth, 2015; Solo, 2016), and the GC F-statistic (Hamilton, 1994, Chap. 11). Finally, we discuss approaches to approximating the distribution of the SR estimator under the alternative hypothesis.

2 Preliminaries

Notation: Throughout, superscript "$\intercal$" denotes matrix transpose and superscript "$*$" conjugate transpose, while superscript "R" refers to a reduced model (Section 2.4); $|\cdots|$ denotes the determinant of a a square matrix and trace$[\cdots]$ its trace; $\|\cdots\|$ denotes a consistent vector/matrix norm; $\mathbf{P}[\cdots]$ denotes probability, $\mathbf{E}[\cdots]$ expectation and var$[\cdots]$ variance; $\xrightarrow{p}$ denotes convergence in probability and $\xrightarrow{d}$ convergence in distribution; "log" denotes natural logarithm. Unless stated otherwise, all vectors are taken to be *column* vectors.

2.1 Maximum likelihood estimation and the large-sample theory

Suppose given a parametrised set of models on a state space $\mathcal{S} = \mathbb{R}^n$, with multivariate parameter $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_r]^\intercal$ in an r-dimensional parameter space $\Theta \subseteq \mathbb{R}^r$. A data sample of size N for the model is then a sequence $\boldsymbol{u} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N\} \in \mathcal{S}^N$; in general, the $\boldsymbol{u}_k$ will not be independent, and we denote the joint distribution of data samples of size N drawn from the model with parameter $\boldsymbol{\theta}$ by $\mathcal{X}_N(\boldsymbol{\theta})$, with probability density function (PDF) $p(\boldsymbol{u}; \boldsymbol{\theta})$. For sample data $\boldsymbol{u} \in \mathcal{S}^N$, the *average log-likelihood function* is defined to be[2]

$$\hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u}) = \frac{1}{N} \log p(\boldsymbol{u}; \boldsymbol{\theta}) \tag{1}$$

i.e., the logarithm of the PDF scaled by sample size. The *maximum likelihood* and *maximum-likelihood parameter estimate* are given respectively by

$$\hat{\ell}(\boldsymbol{u}) = \sup \left\{ \hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u}) : \boldsymbol{\theta} \in \Theta \right\} \tag{2}$$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{u}) \equiv \arg\max \left\{ \hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u}) : \boldsymbol{\theta} \in \Theta \right\} \tag{3}$$

We assume the model is identifiable, so that $\hat{\boldsymbol{\theta}}(\boldsymbol{u})$ is uniquely defined and $\hat{\ell}(\boldsymbol{u}) = \hat{\ell}\big(\hat{\boldsymbol{\theta}}(\boldsymbol{u})|\boldsymbol{u}\big)$. On a point of notation, for any sample statistic $\hat{f}(\boldsymbol{u})$, given $\boldsymbol{\theta} \in \Theta$ we write $\hat{f}(\boldsymbol{\theta})$ for the random variable $\hat{f}(\boldsymbol{U})$ with

[2] We work exclusively with average log-likelihoods, so from here on we drop the "average log-".

$\boldsymbol{U} \sim \mathcal{X}_N(\boldsymbol{\theta})$; $\hat{f}(\boldsymbol{\theta})$ is thus a random variable parametrised by $\boldsymbol{\theta}$. For a parameter $\boldsymbol{\theta}$ itself, we write just $\hat{\boldsymbol{\theta}}$ for its ML estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{U})$.

Let

$$\hat{\ell}_\alpha(\boldsymbol{\theta}|\boldsymbol{u}) = \frac{\partial}{\partial\theta_\alpha}\hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u})\,, \qquad \hat{\ell}_{\alpha\beta}(\boldsymbol{\theta}|\boldsymbol{u}) = \frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u})\,, \qquad \alpha,\beta = 1,\ldots,r \tag{4}$$

The *Fisher information matrix* associated with a parameter $\boldsymbol{\theta} \in \Theta$ is defined as

$$\mathcal{I}_{\alpha\beta}(\boldsymbol{\theta}) \equiv -\mathbf{E}\Big[\hat{\ell}_{\alpha\beta}(\boldsymbol{\theta}|\boldsymbol{U})\Big] = -\int \hat{\ell}_{\alpha\beta}(\boldsymbol{\theta}|\boldsymbol{u})p(\boldsymbol{u};\boldsymbol{\theta})\,d\boldsymbol{u}_1\ldots d\boldsymbol{u}_N \tag{5}$$

i.e., expectation is over $\boldsymbol{U} \sim \mathcal{X}_N(\boldsymbol{\theta})$. Writing $\Omega(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$ for the inverse Fisher information matrix, under certain conditions—which will apply in our case—we have $\forall\boldsymbol{\theta} \in \Theta$

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta} \tag{6}$$

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \Omega(\boldsymbol{\theta})\big) \tag{7}$$

as sample size $N \to \infty$; that is, the maximum-likelihood parameter estimate is consistent and (appropriately scaled) asymptotically normally distributed with mean equal to the true parameter, and covariance given by the inverse Fisher information matrix.

Suppose now that we have a composite nested null hypothesis H_0 defined by $\boldsymbol{\theta} \in \Theta_0$, where $\Theta_0 \subset \Theta$ is the s-dimensional *null subspace*, $s < r$. We write the maximum likelihood and maximum-likelihood parameter under H_0 as

$$\hat{\ell}_0(\boldsymbol{u}) = \sup\left\{\hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u}) : \boldsymbol{\theta} \in \Theta_0\right\} \tag{8}$$

$$\hat{\boldsymbol{\theta}}_0(\boldsymbol{u}) \equiv \arg\max\left\{\hat{\ell}(\boldsymbol{\theta}|\boldsymbol{u}) : \boldsymbol{\theta} \in \Theta_0\right\} \tag{9}$$

respectively. The *(log-)likelihood ratio* statistic given the data sample $\boldsymbol{u}$ is then

$$\hat{\lambda}(\boldsymbol{u}) \equiv 2\left[\hat{\ell}(\boldsymbol{u}) - \hat{\ell}_0(\boldsymbol{u})\right] \tag{10}$$

Note that $\hat{\ell}(\boldsymbol{u}) \geq \hat{\ell}_0(\boldsymbol{u})$ always, so $\hat{\lambda}(\boldsymbol{u}) \geq 0$. Given $\boldsymbol{\theta}$, we write, as described earlier, $\hat{\lambda}(\boldsymbol{\theta})$ for the $\boldsymbol{\theta}$-parametrised random variable $\hat{\lambda}(\boldsymbol{U})$ with $\boldsymbol{U} \sim \mathcal{X}_N(\boldsymbol{\theta})$. Wilks' Theorem (Wilks, 1938) then states that under H_0, the distribution of the sample log-likelihood ratio $\hat{\lambda}(\boldsymbol{\theta})$ is asymptotically χ^2-distributed:

$$\forall\boldsymbol{\theta} \in \Theta_0\,, \quad N\hat{\lambda}(\boldsymbol{\theta}) \xrightarrow{d} \chi^2(d) \tag{11}$$

as sample size $N \to \infty$, where degrees of freedom $d = r - s$ is the difference in dimension of the full and null parameter spaces. Convergence is of order $N^{-\frac{1}{2}}$. Crucially, (11) holds for *any* $\boldsymbol{\theta} \in \Theta_0$; i.e., given that the null hypothesis holds, it doesn't matter *where* in the null space the true parameter lies.

2.2 The generalised χ^2 family of distributions

We introduce a family of distributions that will play a critical role in what follows. Let $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, B)$ be a zero-mean n-dimensional multivariate-normal random vector with covariance matrix B, and A an $n \times n$ symmetric matrix. Then (Jones, 1983) we write $\chi^2(A, B)$ for the distribution of the random quadratic form $Q = \boldsymbol{Z}^\mathsf{T}A\boldsymbol{Z}$. If $A = B = I$, then $\chi^2(A, B)$ reduces to the usual $\chi^2(n)$. If A is $m \times m$ and C is $m \times n$, then $\chi^2(A, CBC^\mathsf{T}) = \chi^2(C^\mathsf{T}AC, B)$.

It is not hard to show (Mohsenipour, 2012) that if B is positive-definite and A symmetric (which will be the case for the generalised χ^2 distributions we encounter), then $\chi^2(A, B) = \chi^2(\Lambda, I)$, where

$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1, \ldots, \lambda_n$ the eigenvalues of BA, or, equivalently, of RAR^T where R is the right-Cholesky factor of B (so that $R^\mathsf{T}R = B$). In that case, we have

$$\lambda_1 U_1^2 + \ldots + \lambda_n U_n^2 \sim \chi^2(A, B)\,, \qquad \text{where } U_i \text{ iid} \sim \mathcal{N}(0,1) \tag{12}$$

so that $\chi^2(A, B)$ is a weighted sum of independent χ^2-distributed variables, and in particular if the λ_i are all equal then we have a scaled $\chi^2(n)$ distribution. From (12), moments of a generalised χ^2 variable may be conveniently expressed in terms of the eigenvalues; thus we may calculate that for $Q \sim \chi^2(A, B)$

$$\mathbf{E}[Q] = \mu \;\; = \;\; \sum_{i=1}^{n} \lambda_i \tag{13a}$$

$$\mathrm{var}[Q] = \sigma^2 = 2\sum_{i=1}^{n} \lambda_i^2 \tag{13b}$$

Empirically, it is found that generalised χ^2 variables (at least for A symmetric and B positive-definite) are very well approximated by Γ distributions: specifically, we have $Q \approx \Gamma(\alpha, \beta)$ with

$$\alpha = \frac{\mu^2}{\sigma^2} \qquad \text{(shape parameter)} \tag{14a}$$

$$\beta = \frac{\sigma^2}{\mu} \qquad \text{(scale parameter)} \tag{14b}$$

2.3 VAR modelling

Given a wide-sense stationary, purely nondeterministic n-dimensional vector stochastic process $\boldsymbol{U}_t = [U_{1t}, \ldots, U_{nt}]^\mathsf{T}$, $-\infty < t < \infty$ (Doob, 1953), under certain conditions (see below) it will have a stable, invertible—and in general infinite-order—VAR representation

$$\boldsymbol{U}_t = \sum_{k=1}^{\infty} A_k \boldsymbol{U}_{t-k} + \boldsymbol{\varepsilon}_t \tag{15}$$

where $\boldsymbol{\varepsilon}_t$ is a white noise process, the sequence of $n \times n$ autoregression coefficient matrices A_k is square-summable, and the $n \times n$ residuals covariance matrix $\Sigma = \mathbf{E}\left[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\mathsf{T}\right]$ is positive-definite. Sufficient conditions for existence of such an invertible VAR representation—and which, importantly, also guarantee a similar representation for any *sub*process—are given in Geweke (1982) (see also Masani, 1966; Rozanov, 1967); we assume those conditions for all vector stochastic processes from now on.

If $A_k = 0$ for $k > p$ then (15) defines a finite-order VAR(p) model. We write $A = [A_1 \ldots A_p]$ (an $n \times pn$ matrix), and the model parameters are given by $\boldsymbol{\theta} = (A, \Sigma)$. The dimensionality of the parameter space is thus $pn^2 + \frac{1}{2}n(n+1)$. The autocovariance sequence for the process $\boldsymbol{U}_t$ is defined by

$$\Gamma_k = \mathbf{E}\left[\boldsymbol{U}_t \boldsymbol{U}_{t-k}^\mathsf{T}\right]\,, \qquad -\infty < k < \infty\,, \tag{16}$$

and we have $\Gamma_{-k} = \Gamma_k^\mathsf{T}$. By a standard trick, the process $\boldsymbol{U}_t^{(p)} = \left[\boldsymbol{U}_t^\mathsf{T}\ \boldsymbol{U}_{t-1}^\mathsf{T} \ldots \boldsymbol{U}_{t-(p-1)}^\mathsf{T}\right]^\mathsf{T}$ satisfies a pn-dimensional VAR(1) model

$$\boldsymbol{U}_t^{(p)} = \boldsymbol{A}\boldsymbol{U}_{t-1}^{(p)} + \boldsymbol{\varepsilon}_t^{(p)} \tag{17}$$

where the $pn \times pn$ "companion matrix" is given by

$$\boldsymbol{A} = \begin{bmatrix} A_1 & A_2 & \ldots & A_{p-1} & A_p \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix} \tag{18}$$

and $\boldsymbol{\varepsilon}_t^{(p)} = \left[\boldsymbol{\varepsilon}_t^\mathsf{T}\ 0 \dots 0\right]^\mathsf{T}$ with $pn \times pn$ covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \tag{19}$$

The *spectral radius* of the model is given by the largest absolute eigenvalue of $\boldsymbol{A}$:

$$\rho(A) = \max\{|z| : |Iz - \boldsymbol{A}| = 0\} \tag{20}$$

The model is stable iff $\rho(A) < 1$.

Taking the covariance of both sides of (17) yields

$$\boldsymbol{\Gamma} - \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^\mathsf{T} = \boldsymbol{\Sigma} \tag{21}$$

where $\boldsymbol{\Gamma}$ is the $pn \times pn$ covariance matrix

$$\boldsymbol{\Gamma} = \mathbf{E}\left[\boldsymbol{U}_t^{(p)}\boldsymbol{U}_t^{(p)\mathsf{T}}\right] \tag{22}$$

The $k\ell$-block of $\boldsymbol{\Gamma}$ is given by $\boldsymbol{\Gamma}_{k\ell} = \Gamma_{\ell-k}$ for $k, \ell = 1, \dots, p$. (21) is a discrete-time Lyapunov (DLYAP) equation—which may be readily solved numerically—and is equivalent to (the first p of) the Yule-Walker equations

$$\Gamma_k = \sum_{\ell=1}^{p} A_\ell \Gamma_{k-\ell} + \delta_{k0}\Sigma \qquad -\infty < k < \infty \tag{23}$$

If the parameters (A, Σ) are known, the autocovariance sequence may be calculated from (21), or recursively from (23); conversely, (A, Σ) may be calculated from $\Gamma_0, \dots, \Gamma_p$, e.g., by Whittle's algorithm (Whittle, 1963). In sample, given a data sequence $\boldsymbol{u} = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_N\}$, ML parameters $\left(\hat{A}(\boldsymbol{u}), \hat{\Sigma}(\boldsymbol{u})\right)$ may be calculated via a standard ordinary least squares (OLS).

In the spectral domain (Hannan and Deistler, 2012), let $\omega \in [0, 2\pi]$ denote angular frequency in radians. The *transfer function* for the VAR model (15) is defined as

$$\Psi(\omega) = \Phi(\omega)^{-1} \tag{24}$$

where

$$\Phi(\omega) = I - \sum_{k=1}^{\infty} A_k e^{-i\omega k} \tag{25}$$

is the Fourier transform of the VAR coefficients sequence. The *cross-power spectral density* (CPSD) matrix $S(\omega)$ is given by the Fourier transform of the autocovariance sequence

$$S(\omega) = \sum_{k=-\infty}^{\infty} \Gamma_k e^{-i\omega k} \tag{26}$$

and conversely, the autocovariance sequence is the inverse transform of the CPSD:

$$\Gamma_k = \frac{1}{2\pi}\int_0^{2\pi} S(\omega) e^{i\omega k} d\omega \tag{27}$$

$S(\omega)$ is Hermitian $\forall\omega$, and satisfies the factorisation (Wiener and Masani, 1957)

$$S(\omega) = \Psi(\omega)\Sigma\Psi(\omega)^* \tag{28}$$

The CPSD $S(\omega)$, via (28), uniquely determines the VAR parameters, which may be factored out computationally, e.g., by Wilson's algorithm (Wilson, 1972).

2.4 Granger-Geweke causality

Geweke (1982) defines the population (unconditional[3]) Granger causality statistic in the following context: suppose that the process (15) is partitioned into subprocesses $\boldsymbol{U}_t = [\boldsymbol{X}_t^\mathsf{T}\ \boldsymbol{Y}_t^\mathsf{T}]^\mathsf{T}$ of dimension n_x, n_y respectively. The assumed regularity conditions on $\boldsymbol{U}_t$ (Geweke, 1982, Sec. 2) ensure that the subprocess $\boldsymbol{X}_t$ will itself admit a stable, invertible VAR representation

$$\boldsymbol{X}_t = \sum_{k=1}^{\infty} A_k^\mathsf{R} \boldsymbol{X}_{t-k} + \boldsymbol{\varepsilon}_t^\mathsf{R} \tag{29}$$

with square-summable coefficients A_k^R and positive-definite residuals covariance matrix $\Sigma^\mathsf{R} = \mathbf{E}\big[\boldsymbol{\varepsilon}_t^\mathsf{R}\boldsymbol{\varepsilon}_t^{\mathsf{R}\mathsf{T}}\big]$. To define Granger causality, the *reduced* regression (29) is contrasted with the *full* regression; that is, the x-component

$$\boldsymbol{X}_t = \sum_{k=1}^{\infty} A_{k,xx} \boldsymbol{X}_{t-k} + \sum_{k=1}^{\infty} A_{k,xy} \boldsymbol{Y}_{t-k} + \boldsymbol{\varepsilon}_{xt} \tag{30}$$

of (15). We stress here that:

1. The reduced model parameters $(A^\mathsf{R}, \Sigma^\mathsf{R})$ are fully determined by the full model parameters (A, Σ).
2. Even if the full regression (30) has finite order, the reduced regression (29) will in general *not* have finite order.

The population Granger causality from $\boldsymbol{Y} \to \boldsymbol{X}$ for the VAR (15) with parameters $\boldsymbol{\theta}$ is then defined as

$$F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) = \log \frac{\left|\Sigma^\mathsf{R}\right|}{|\Sigma_{xx}|} \tag{31}$$

The justification for the description of (31) as a "causal" statistic, is as follows: $\boldsymbol{\varepsilon}_{xt}$ in (30) represents the residual error associated with the optimal least-squares prediction $\mathbf{E}[\boldsymbol{X}_t \,|\, \boldsymbol{U}_{t-1}, \boldsymbol{U}_{t-2}, \ldots] = \sum_{k=1}^{\infty} A_{k,xx}\boldsymbol{X}_{t-k} + \sum_{k=1}^{\infty} A_{k,xy}\boldsymbol{Y}_{t-k}$ of $\boldsymbol{X}_t$ by its own past $\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots$ and the past $\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-2}, \ldots$ of $\boldsymbol{Y}_t$. By contrast, $\boldsymbol{\varepsilon}_t^\mathsf{R}$ in (29) represents the residual error associated with the optimal prediction $\mathbf{E}[\boldsymbol{X}_t \,|\, \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots] = \sum_{k=1}^{\infty} A_k^\mathsf{R}\boldsymbol{X}_{t-k}$ of $\boldsymbol{X}_t$ by its own past alone. Magnitudes of residual prediction errors are then quantified by the generalised variances $|\Sigma_{xx}|$ and $\left|\Sigma^\mathsf{R}\right|$ respectively (Wilks, 1932; Barrett et al., 2010), leading to the interpretation of (31) as the extent to which inclusion of the past $\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-2}, \ldots$ of $\boldsymbol{Y}_t$ in the predictor set reduces the prediction error for $\boldsymbol{X}_t$.

In the frequency domain, Geweke (1982) defines the (population, unconditional) spectral Granger causality at angular frequency ω by

$$f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta}) = \log \frac{|S_{xx}(\omega)|}{\left|S_{xx}(\omega) - \Psi_{xy}(\omega)\Sigma_{yy|x}\Psi_{xy}(\omega)^*\right|} \tag{32}$$

Barnett and Seth (2011) introduce *band-limited* (frequency-averaged) spectral Granger causality

$$f_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta}) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})\, d\omega \tag{33}$$

where the frequency range $\mathcal{F}$ is a measurable subset of $[0, 2\pi]$ (in practice usually an interval). $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ averaged across all frequencies yields the corresponding time-domain GC (Geweke, 1982); that is,

$$f_{\boldsymbol{Y}\to\boldsymbol{X}}([0, 2\pi];\boldsymbol{\theta}) = \frac{1}{2\pi}\int_0^{2\pi} f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})\, d\omega = F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) \tag{34}$$

[3]In this study we only address the unconditional case; but see Section 5.2.

2.4.1 Likelihood-ratio estimation

Suppose given a finite-order VAR model.

$$\boldsymbol{U}_t = \sum_{k=1}^{p} A_k \boldsymbol{U}_{t-k} + \boldsymbol{\varepsilon}_t \tag{35}$$

for the process $\boldsymbol{U}_t = [\boldsymbol{X}_t^\mathsf{T} \ \boldsymbol{Y}_t^\mathsf{T}]^\mathsf{T}$. For now, we assume that the model order p is known (in Section 5.1 we discuss infinite-order VAR models and model order selection). How then, given a data sequence $\boldsymbol{u} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$ sampled from (35) with unknown parameters $\boldsymbol{\theta} = (A, \Sigma)$, might $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ be *estimated*? On the face of it, $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ is—uncoincidentally, as remarked in Geweke (1982)—a population log-likelihood-ratio statistic, since the maximum (average) log-likelihood for a finite-order VAR model of the form (35) is, up to a constant additive term, just $-\frac{1}{2}\log|\hat{\Sigma}(\boldsymbol{u})|$, where $\hat{\Sigma}(\boldsymbol{u})$ is the ML (OLS) estimate for the actual residuals covariance matrix Σ. But as regards estimation of $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$, a problem arises: while the full model order p may be finite, the reduced model (29) will in general be of *infinite* order. We might be tempted—as suggested in Geweke (1982, 1984), and, until recently, standard practice—to simply truncate the reduced model at order p. Then (29) becomes a nested sub-model of (30), and plugging in maximum-likelihood estimates $\hat{\Sigma}(\boldsymbol{u})$, $\hat{\Sigma}^\mathsf{R}(\boldsymbol{u})$ for the residuals covariance matrices Σ, Σ^R in (31) yields a true log-likelihood-ratio sample statistic

$$\hat{F}^{\mathsf{LR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{u}) = \log \frac{\left|\hat{\Sigma}^\mathsf{R}(\boldsymbol{u})\right|}{|\hat{\Sigma}_{xx}(\boldsymbol{u})|} \tag{36}$$

for the composite null hypothesis

$$H_0 : A_{1,xy} = \ldots = A_{p,xy} = 0 \tag{37}$$

But here a problem arises: the resulting reduced model will be misspecified, and failure to take into account sufficient lags of $\boldsymbol{X}_t$ in the reduced regression biases the resulting Granger causality estimator. Noting that a VAR(p) model is also VAR(q) for $q > p$ (coefficients $A_k, p < k \leq q$ can simply be set to zero), we could attempt to remedy the situation by selecting a parsimonious model order $q > p$ for the reduced model by a standard (data sample length-dependent) model order selection criterion (McQuarrie and Tsai, 1998), and extend the full model to order q. However, in doing so the full model becomes over-specified and the variance of the resulting estimator is inflated. Furthermore, since the estimated model order will increase with sample length N, it is not clear whether the estimator will be consistent in any meaningful sense. We discuss this further in Section 5.1. This conundrum was explicitly identified by Stokes and Purdon (2017)[4], although its symptoms had previously been noted, particularly in the spectral domain (see e.g., Ding et al., 2006; Chen et al., 2006)[5].

2.4.2 Single-regression estimation

The above issues may, however, be sidestepped. Given a finite-order VAR(p) model (35) (again, we assume that p is known, and discuss infinite-order VAR models and model order selection in Section 5.1), the reduced VAR (29) may not be assumed finite-dimensional, but the reduced residuals covariance matrix Σ^R will nonetheless be a continuous, *deterministic* function

$$\Sigma^\mathsf{R} = V(\boldsymbol{\theta}) \tag{38}$$

[4] Stokes and Purdon (2017), having identified the LR estimator as problematic, concede that at the time they were unaware that there were already estimators which obviate the problem (Stokes and Purdon, 2018). Other claims made in Stokes and Purdon (2017) have also come under critical scrutiny (Faes et al., 2017; Barnett et al., 2018a,b; Dhamala et al., 2018).

[5] But note that the "block-decomposition" method presented in Chen et al. (2006) to address the conundrum—essentially an attempt at constructing a single-regression estimator (see Section 2.4.2 below)—is incorrect (Solo, 2016).

of the finite-dimensional full-model parameters $\boldsymbol{\theta} = (A, \Sigma)$, with $V(\boldsymbol{\theta}) = \Sigma_{xx}$ for $\boldsymbol{\theta} \in \Theta_0$. Given parameters (A, Σ), the function $V(\boldsymbol{\theta})$ may be computed numerically to desired precision by spectral factorisation in the frequency domain (Dhamala et al., 2008a,b), spectral factorisation in the time domain domain (Barnett and Seth, 2014) or by a state-space method (Barnett and Seth, 2015; Solo, 2016) which devolves to solution of a discrete algebraic Riccati equation (DARE); see Appendix B. Now from (31) and (38) the population GC is

$$F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) = \log \frac{|V(\boldsymbol{\theta})|}{|\Sigma_{xx}|} \tag{39}$$

Then given a data sample $\boldsymbol{u} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$ we need only estimate the full VAR(p) model (35), to obtain full-model ML parameter estimates $\hat{\boldsymbol{\theta}}(\boldsymbol{u})$; the reduced model estimate $\hat{\Sigma}^{\mathrm{R}}(\boldsymbol{u}) = V\big(\hat{\boldsymbol{\theta}}(\boldsymbol{u})\big)$ is calculated directly from the full-model estimates by one of the techniques described above, yielding the *single-regression Granger causality estimator*

$$\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{u}) = \log \frac{\big|V\big(\hat{\boldsymbol{\theta}}(\boldsymbol{u})\big)\big|}{\big|\hat{\Sigma}_{xx}(\boldsymbol{u})\big|} \tag{40}$$

Since ML parameter estimates are consistent, $\hat{\Sigma}_{xx}\big(\hat{\boldsymbol{\theta}}\big) \xrightarrow{p} \Sigma_{xx}$, and by the Continuous Mapping Theorem (CMT; Lehmann and Romano, 2005) $V\big(\hat{\boldsymbol{\theta}}\big) \xrightarrow{p} \Sigma^{\mathrm{R}}$. Thus the SR estimator $\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) = F_{\boldsymbol{Y}\to\boldsymbol{X}}\big(\hat{\boldsymbol{\theta}}\big)$ is consistent.

3 Asymptotic null distribution for single-regression GC estimators

As regards statistical inference, LR sample statistics in general satisfy Wilks' Theorem (Wilks, 1938), which implies that they are asymptotically $\chi^2(d)$-distributed under the null hypothesis as sample size $N \to \infty$, with degrees of freedom $d = qn_xn_y$, where q is the model order used for the full and reduced regressions; see Section 5.1 for further discussion.

The SR estimator $\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$, however, is *not* a log-likelihood ratio, so Wilks' Theorem does not apply, and the asymptotic distribution under the null hypothesis (37) has thus far remained unknown[6]. This is the principal subject of our study. We shall see that, unlike Wilks' asymptotic null χ^2 distribution, the sampling distribution of the single-regression estimator under the null depends explicitly on the (true) null parameters $\boldsymbol{\theta} \in \Theta_0$ themselves, raising some awkward questions regarding statistical inference, which we address in Section 4.

We proceed with a technical result—essentially a multivariate 2nd-order Delta Method (Lehmann and Romano, 2005)—on which our derivation of the asymptotic distributions for the time-domain and band-limited SR estimators hinges:

Proposition 3.1. Let $f(\boldsymbol{\theta})$ be a non-negative, twice-differentiable function on a smooth r-dimensional manifold Θ which vanishes identically on the s-dimensional hyperplane[7] $\Theta_0 \subset \Theta$ specified by $\theta_1 = \ldots = \theta_d = 0$ with $d = r - s$. Then

a. The gradient $\nabla f(\boldsymbol{\theta})$ is zero for all $\boldsymbol{\theta} \in \Theta_0$.

b. Writing a subscript "$_0$" to denote the $d \times d$ upper-left submatrix of an $r \times r$ matrix, for $\boldsymbol{\theta} \in \Theta_0$ the

[6]The VAR single-regression estimator is a special case of the more general state-space GC estimator (Barnett and Seth, 2015; Solo, 2016). Solo (2016) claims without proof (and, as we shall see, incorrectly), that the asymptotic distribution of this estimator will in general be χ^2 under the null hypothesis. Barnett and Seth (2015), find through extensive simulation that the state-space estimator under the null is well-approximated by a Γ distribution; this, we shall see, is well-explained here, at least in the VAR case.

[7]If Θ_0 is a general smooth submanifold of Θ, then Proposition 3.1 still holds around any given $\boldsymbol{\theta} \in \Theta_0$ modulo an appropriate local transformation of coordinates.

Hessian $\mathbf{H}(\boldsymbol{\theta}) = \nabla^2 f(\boldsymbol{\theta})$ takes the form

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{H}_0(\boldsymbol{\theta}) & 0 \\ 0 & 0 \end{bmatrix} \tag{41}$$

with $\mathbf{H}_0(\boldsymbol{\theta})$ positive-semidefinite.

c. For $\boldsymbol{\theta} \in \Theta_0$ let $\boldsymbol{\vartheta}_N$ be a sequence of r-dimensional random vectors with $\sqrt{N}(\boldsymbol{\vartheta}_N - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Omega)$ as $N \to \infty$; *cf.* (7). Then

$$N f(\boldsymbol{\vartheta}_N) \xrightarrow{d} \chi^2\left(\tfrac{1}{2}\mathbf{H}_0(\boldsymbol{\theta}), \Omega_0\right) \text{ as } N \to \infty \tag{42}$$

See Appendix A for a proof.

3.1 The time-domain SR estimator

We shall apply Proposition 3.1 with $f(\boldsymbol{\theta})$ given by the $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ of (39). Firstly, $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ is non-negative and vanishes on Θ_0 (Geweke, 1982). Below we establish that it is also twice-differentiable (in fact *analytic*), so that by Proposition 3.1 with $\boldsymbol{\vartheta}_N$ the ML parameter estimate $\hat{\boldsymbol{\theta}}$ for sample size N, we have for any $\boldsymbol{\theta} \in \Theta_0$

$$N \hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) \xrightarrow{d} \chi^2\left(\tfrac{1}{2}\mathbf{H}_0(\boldsymbol{\theta}), \Omega_0(\boldsymbol{\theta})\right) \text{ as } N \to \infty \tag{43}$$

where $\mathbf{H}(\boldsymbol{\theta})$ is the Hessian of $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$ the inverse Fisher information matrix for the VAR(p) model (35) evaluated at $\boldsymbol{\theta}$. Here the "$_0$" subscript denotes a submatrix corresponding to the null-hypothesis variable indices $x = \{1, \ldots, n_x\}$, $y = \{n_x + 1, \ldots, n\}$.

To calculate the generalised χ^2 parameters, we thus require firstly the null submatrix $\Omega_0(\boldsymbol{\theta})$ of the inverse Fisher information matrix $\Omega(\boldsymbol{\theta})$ for a VAR model specified by $\boldsymbol{\theta} = (A, \Sigma)$; this is a standard result. Let $\boldsymbol{\Gamma}$ be the autocovariance matrix of (22). Considering multi-indices $[k, ij]$ for the regression coefficients $A_{k,ij}$ (so that k indexes lags and i, j variables), the entries for the inverse Fisher information matrix corresponding to the $A_{k,ij}$ are given by (Hamilton, 1994; Lütkepohl, 1993)

$$\Omega(\boldsymbol{\theta})_{[k,ij][k',i'j']} = \Sigma_{ii'}\left[\boldsymbol{\Gamma}^{-1}\right]_{kk',jj'} \tag{44}$$

where $\left[\boldsymbol{\Gamma}^{-1}\right]_{kk',jj'}$ denotes the jj' entry of the kk'-block of $\boldsymbol{\Gamma}^{-1}$; or, expressed as a Kronecker product:

$$\text{autoregression coefficients block of } \Omega(\boldsymbol{\theta}) \;= \Sigma \otimes \boldsymbol{\Gamma}^{-1} \tag{45}$$

$\Omega_0(\boldsymbol{\theta})$ is then given by the submatrix of (44) with $i, i' \in x$ and $j, j' \in y$, which we write as

$$\Omega_0(\boldsymbol{\theta}) = \Sigma_{xx} \otimes \left[\boldsymbol{\Gamma}^{-1}\right]_{yy} \tag{46}$$

Secondly, to calculate the null Hessian $\mathbf{H}_0(\boldsymbol{\theta})$, we require an expression for the function $V(\boldsymbol{\theta})$ of (38). This we accomplish via the state-space formalism introduced in Barnett and Seth (2015). In Appendix B we show that

$$V(\boldsymbol{\theta}) = A_{xy}\boldsymbol{P}A_{xy}^{\mathsf{T}} + \Sigma_{xx} \tag{47}$$

where the $pn_y \times pn_y$ symmetric matrix $\boldsymbol{P}$ is the solution of the discrete algebraic Riccati equation (DARE)

$$\boldsymbol{P} - \boldsymbol{A}_{yy}\boldsymbol{P}\boldsymbol{A}_{yy}^{\mathsf{T}} = \boldsymbol{\Sigma}_{yy} - \left(\boldsymbol{A}_{yy}\boldsymbol{P}A_{xy}^{\mathsf{T}} + \boldsymbol{\Sigma}_{yx}\right)\left(A_{xy}\boldsymbol{P}A_{xy}^{\mathsf{T}} + \Sigma_{xx}\right)^{-1}\left(A_{xy}\boldsymbol{P}\boldsymbol{A}_{yy}^{\mathsf{T}} + \boldsymbol{\Sigma}_{yx}^{\mathsf{T}}\right) \tag{48}$$

with

$$\boldsymbol{A}_{yy} = \begin{bmatrix} A_{1,yy} & A_{2,yy} & \dots & A_{p-1,yy} & A_{p,yy} \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix} \qquad A_{xy} = \begin{bmatrix} A_{1,xy} & A_{2,xy} & \dots & A_{p-1,xy} & A_{p,xy} \end{bmatrix} \tag{49}$$

and

$$\boldsymbol{\Sigma}_{yy} = \begin{bmatrix} \Sigma_{yy} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{yx} = \begin{bmatrix} \Sigma_{yx} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{50}$$

It is not hard to see that, by construction, $F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) = \log|V(\boldsymbol{\theta})| - \log|\Sigma_{xx}|$ is an analytic function of $\boldsymbol{\theta}$: calculation of $V(\boldsymbol{\theta})$ via (48) and (47) only involves algebraic operations (solution of multivariate polynomial equations), and we know $V(\boldsymbol{\theta})$ to be positive-definite, so that $|V(\boldsymbol{\theta})| > 0\ \forall\boldsymbol{\theta}$ and $\log|V(\boldsymbol{\theta})|$ is thus analytic.

To calculate $\mathbf{H}_0(\boldsymbol{\theta})$ we require derivatives up to 2nd order of $V(\boldsymbol{\theta})$, as defined implicitly through (47) and (48), with respect to the null-hypothesis parameters (that is, with respect to $A_{k,ij}$ for $i \in x$, $j \in y$), evaluated for $\boldsymbol{\theta} \in \Theta_0$. From (47) we may calculate:

$$\frac{\partial V_{ii'}}{\partial A_{k,uv}} = \delta_{ui}\left[\boldsymbol{P}A_{xy}^{\mathsf{T}}\right]_{k,vi'} + \delta_{ui'}\left[A_{xy}\boldsymbol{P}\right]_{k,iv} + \left[A_{xy}\frac{\partial \boldsymbol{P}}{\partial A_{k,uv}}A_{xy}^{\mathsf{T}}\right]_{ii'} \tag{51}$$

where indices $i, i', u, u' \in x$, indices $j, j', v, v' \in y$ and $k, k' = 1, \dots, p$. Since A_{xy} vanishes under the null hypothesis, we have

$$\left.\frac{\partial V_{ii'}}{\partial A_{k,uv}}\right|_{\boldsymbol{\theta}\in\Theta_0} = 0 \tag{52}$$

and from (51) we find

$$\left.\frac{\partial^2 V_{ii'}}{\partial A_{k,uv}\partial A_{k',u'v'}}\right|_{\boldsymbol{\theta}\in\Theta_0} = \left[\delta_{ui}\delta_{u'i'} + \delta_{ui'}\delta_{u'i}\right]\boldsymbol{P}_{kk',vv'} \tag{53}$$

We see then that $\boldsymbol{P}$ is required only on the null space $A_{xy} = 0$, in which case the DARE (48) becomes a DLYAP equation:

$$\boldsymbol{P} - \boldsymbol{A}_{yy}\boldsymbol{P}\boldsymbol{A}_{yy}^{\mathsf{T}} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}[\Sigma_{xx}]^{-1}\boldsymbol{\Sigma}_{yx}^{\mathsf{T}} \tag{54}$$

We may now calculate the required Hessian. For null parameters $\theta_\alpha, \theta_\beta$, from the definition (39) and using (52) we may calculate

$$\left.\frac{\partial^2 F_{\boldsymbol{Y}\to\boldsymbol{X}}}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\boldsymbol{\theta}\in\Theta_0} = \left.\frac{\partial^2 \log|V|}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\boldsymbol{\theta}\in\Theta_0} = \operatorname{trace}\left[[\Sigma_{xx}]^{-1}\left.\frac{\partial^2 V}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\boldsymbol{\theta}\in\Theta_0}\right] \tag{55}$$

(53) then yields

$$[\mathbf{H}_0(\boldsymbol{\theta})]_{[k,uv],[k',u'v']} = 2\left[[\Sigma_{xx}]^{-1}\right]_{uu'}\boldsymbol{P}_{kk',vv'} \tag{56}$$

or

$$\tfrac{1}{2}\mathbf{H}_0(\boldsymbol{\theta}) = [\Sigma_{xx}]^{-1} \otimes \boldsymbol{P} \tag{57}$$

We note that in (54),

$$\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}[\Sigma_{xx}]^{-1}\boldsymbol{\Sigma}_{yx}^{\mathsf{T}} = \boldsymbol{\Sigma}_{yy|x} = \begin{bmatrix} \Sigma_{yy|x} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \tag{58}$$

where $\Sigma_{yy|x} - \Sigma_{yx}[\Sigma_{xx}]^{-1}\Sigma_{xy}$ is a partial covariance matrix. Thus (54) [*cf.* (21)] specifies the autocovariance matrix [*cf.* (22)] for a notional n_y-dimensional VAR(p) model with parameters $(A_{yy}, \Sigma_{yy|x})$. Accordingly, we write $\boldsymbol{P}$ as $\boldsymbol{\Gamma}_{yy|x}$ from now on, so that (57) becomes $\frac{1}{2}\mathbf{H}_0(\boldsymbol{\theta}) = [\Sigma_{xx}]^{-1} \otimes \boldsymbol{\Gamma}_{yy|x}$, and from (43) and (46) it follows that

$$N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) \xrightarrow{d} \chi^2\left([\Sigma_{xx}]^{-1} \otimes \boldsymbol{\Gamma}_{yy|x}, \Sigma_{xx} \otimes \left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\right) \quad \text{as } N \to \infty \tag{59}$$

But from the transformation invariance $\chi^2(CAC^{\mathsf{T}}, B) = \chi^2(A, C^{\mathsf{T}}BC)$ and the mixed-product property of the Kronecker product, we may verify that the Σ_{xx} terms cancel (Σ_{xx} is positive-definite, and thus has an invertible Cholesky decomposition). We thus state our first principal result as

Theorem 3.1. The asymptotic distribution of the single-regression Granger causality estimator under the null hypothesis $\boldsymbol{\theta} \in \Theta_0$ (i.e., $A_{k,xy} = 0\,\forall k$) is given by

$$N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta}) \xrightarrow{d} \chi^2\left(\boldsymbol{I}_{xx} \otimes \boldsymbol{\Gamma}_{yy|x}, \boldsymbol{I}_{xx} \otimes \left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\right) \quad \text{as } N \to \infty \tag{60}$$

where $\boldsymbol{I}_{xx}$ is the $n_x \times n_x$ identity matrix, and $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_{yy|x}$ satisfy the respective DLYAP equations

$$\boldsymbol{\Gamma} - \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{\Sigma} \tag{61a}$$

$$\boldsymbol{\Gamma}_{yy|x} - \boldsymbol{A}_{yy}\boldsymbol{\Gamma}_{yy|x}\boldsymbol{A}_{yy}^{\mathsf{T}} = \boldsymbol{\Sigma}_{yy|x} \tag{61b}$$

or, equivalently, the corresponding Yule-Walker equations (23). ■

From (60) we see that the limiting distribution of $N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ is the sum of n_x random variables independently and identically distributed as $\chi^2\left(\boldsymbol{\Gamma}_{yy|x}, \left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\right)$. Through (12), this distribution may be expressed in terms of the eigenvalues of $\boldsymbol{I}_{xx} \otimes \left(\left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\boldsymbol{\Gamma}_{yy|x}\right)$; these are in fact the eigenvalues $\lambda_1, \ldots, \lambda_{pn_y}$ of $\left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\boldsymbol{\Gamma}_{yy|x}$, where each λ_i appears with multiplicity n_x. The asymptotic distribution of $N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ under the null thus takes the form of a weighted sum of pn_y iid $\chi^2(n_x)$ variables:

$$\lambda_1 W_1 + \ldots + \lambda_{pn_y} W_{pn_y}\,, \qquad W_i \text{ iid} \sim \chi^2(n_x) \tag{62}$$

The asymptotic mean and variance of $N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0$ are

$$\mathbf{E}\left[N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})\right] \to n_x \sum_{i=1}^{pn_y} \lambda_i \tag{63a}$$

$$\operatorname{var}\left[N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})\right] \to 2n_x \sum_{i=1}^{pn_y} \lambda_i^2 \tag{63b}$$

respectively, from which the Γ-approximation of the generalised χ^2 distribution may be obtained, via (13) and (14). Figure 1 plots generalised χ^2, Γ-approximation and empirical SR-estimator cumulative density functions (CDFs) for a representative null VAR model with $n_x = 3$, $n_y = 5$ and $p = 7$ for several sample sizes, illustrating asymptotic convergence with increasing sample length N. The Γ-approximation is barely distinguishable from the generalised χ^2.

We note that the eigenvalues λ_i are all > 0, since both $\left[\boldsymbol{\Gamma}^{-1}\right]_{yy}$ and $\boldsymbol{\Gamma}_{yy|x}$ are guaranteed to be positive-definite (this follows from the purely nondeterministic assumption on the process $\boldsymbol{U}_t$). From (14a) and (63), we find that the shape parameter of the Γ-approximation satisfies[8]

$$\frac{n_x}{2} \le \alpha \le \frac{pn_x n_y}{2}\,, \tag{64}$$

[8]This follows from the identities $\left(\sum_{i=1}^{pn_y}\lambda_i\right)^2 = \sum_{i=1}^{pn_y}\lambda_i^2 + \sum_{i,j=1}^{pn_y}\lambda_i\lambda_j = pn_y\sum_{i=1}^{pn_y}\lambda_i^2 - \frac{1}{2}\sum_{i,j=1}^{pn_y}(\lambda_i - \lambda_j)^2$.

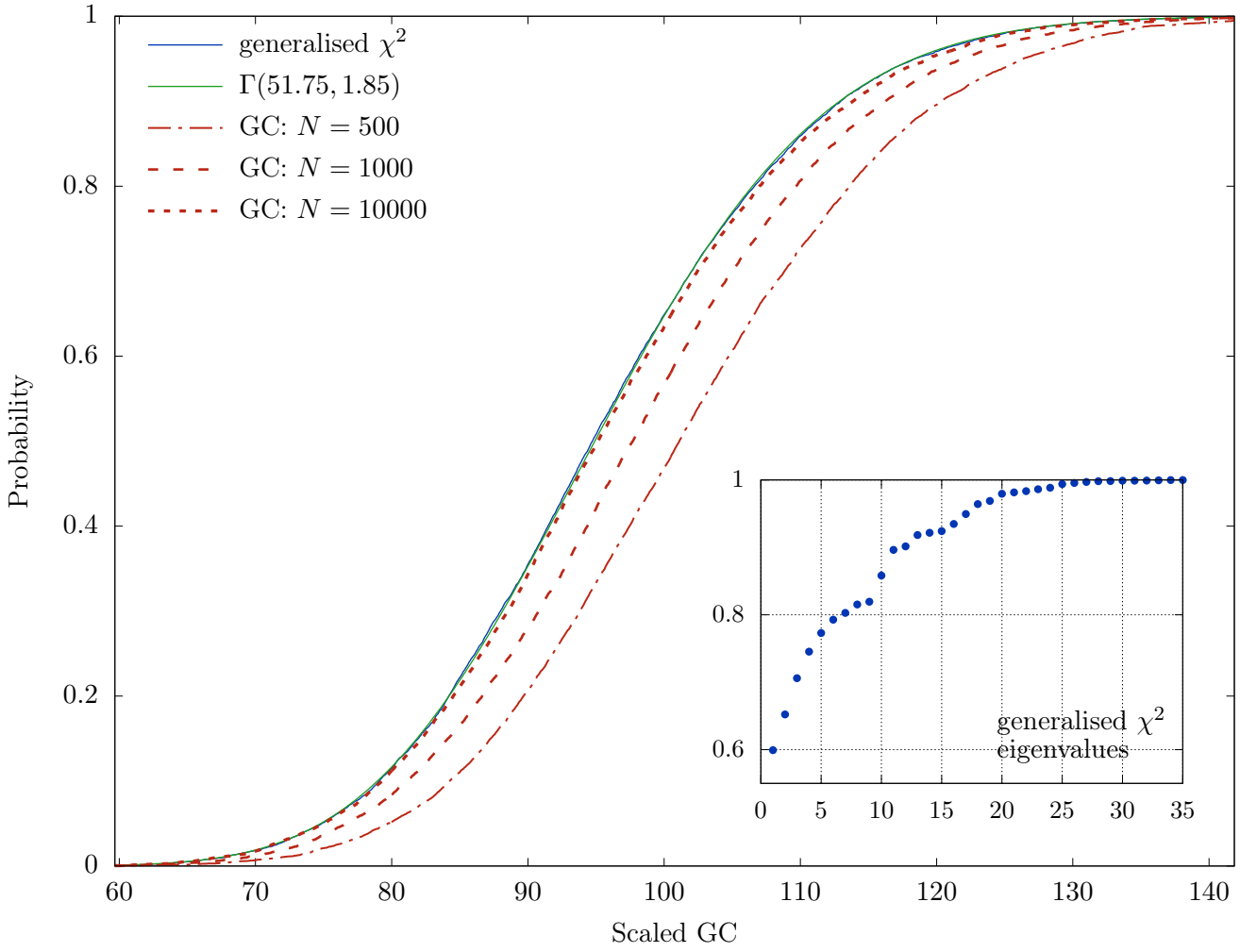

Figure 1: Generalised χ^2, Γ-approximation and empirical SR-estimator CDFs for a representative null VAR model with $n_x = 3$, $n_y = 5$ and $p = 7$, for several sample sizes. The null VAR model was randomly generated according to the scheme described in Appendix D, with spectral radius $\rho = 0.9$ and residuals generalised correlation $\gamma = 1$. The generalised χ^2 and Γ-approximation are almost indistinguishable. Empirical GC plots were based on 10^4 generated time series for each sample lengths N. Inset figure: the $pn_y = 35$ distinct eigenvalues for the generalised χ^2 distribution, sorted by size. (Each eigenvalue will be repeated $n_x = 3$ times.)

and $\alpha = \frac{pn_xn_y}{2} \iff$ all the λ_i are equal $\iff p = n_y = 1$, in which case the distribution of $N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ is asymptotically $\chi^2(n_x)$ scaled by λ (*cf.* Section 3.3). We also state the following conjecture, which we have tested extensively empirically, but have so far been unable to prove rigorously:

Conjecture 3.2. The eigenvalues of $\left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\boldsymbol{\Gamma}_{yy|x}$ satisfy $\lambda_i \leq 1$ for $i = 1, \ldots, pn_y$.

If Conjecture 3.2 holds, then from (14b) and (63) the scale parameter of the Γ-approximation satisfies:

$$0 \leq \beta \leq 2 \tag{65}$$

Simulations reveal that spectral radius and residuals generalised correlation of null parameters $\boldsymbol{\theta}$ have a strong effect on the distribution of the eigenvalues λ_i. Spectral radius close to 1 and large residuals correlation give rise to a larger spread of eigenvalues < 1, resulting in asymptotic null sampling distributions significantly different from a (non-generalised) χ^2. Figure 2 presents the distribution of single-regression estimator CDFs under random sampling of VAR models of given size, spectral radius and residuals generalised correlation (see Appendix D for the VAR sampling method), where the effects of spectral radius and residuals generalised correlation on the null distribution via the eigenvalues (inset figures) is clearly seen.

Assuming Conjecture 3.2, an immediate consequence of (63) and $N\hat{F}^{\mathsf{LR}}_{\boldsymbol{Y}\to\boldsymbol{X}} \xrightarrow{d} \chi^2(pn_xn_y)$ is that for $\boldsymbol{\theta} \in \Theta_0$

$$\mathbf{E}\left[N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})\right] \leq \mathbf{E}\left[N\hat{F}^{\mathsf{LR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})\right] \tag{66a}$$

$$\mathrm{var}\left[N\hat{F}^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})\right] \leq \mathrm{var}\left[N\hat{F}^{\mathsf{LR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})\right] \tag{66b}$$

This suggests that, more generally, the SR estimator will have smaller bias and better efficiency than the LR estimator[9]. As is apparent in Figure 2, the reduction in bias and variance is stronger for spectral radius close to 1 and high residuals correlation.

3.2 The band-limited spectral SR estimator

We now consider the asymptotic null-distribution of the band-limited spectral Granger Causality estimator $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$ (33) which facilitates inference on the corresponding band-limited population statistic $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$. As explained below, the appropriate null hypothesis in this case is in fact identical to the time-domain (or "broadband") null condition H_0. Inference on $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$ is nevertheless informative beyond a time-domain test; thus while we may reject H_0 in the neighbourhood of ω_1 at some significance level, we may fail to reject H_0 at the same level around a different frequency ω_2, with the implication that while H_0 likely does not hold, Granger causality is likely to be significant around ω_1 but negligible around ω_2. In other words, while the time-domain test is sensitive to *any* (sufficiently large) deviation from the null hypothesis—regardless of localisation in the frequency spectrum—the band-limited test is sensitive to discrepancies within the specified frequency band, but insensitive to discrepancies outside this band. Thus a significant band-limited test can justifiably be attributed to a sizeable contribution from the frequency band in question, whereas a significant time-domain test allows only to draw the less-specific conclusion that contributions from across the entire frequency range are potentially implicated in the result.

As we shall see in the following, the limiting asymptotic distribution under H_0 of the band-limited estimator $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}([\omega-\varepsilon, \omega+\varepsilon];\boldsymbol{\theta})$ as $\varepsilon \to 0$ exists. It should *not*, however, be conflated with the asymptotic distribution of $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ under the *point-frequency null hypothesis* $H_0(\omega)$ (see Section 5.2 for discussion of the point-frequency case).

The point-frequency spectral GC $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ is non-negative, and for any ω clearly vanishes under H_0. We note further that by assumed stability of the VAR(p) (35), the inverse transfer function $\Phi(\omega)$ does

[9]Strictly speaking we require the distributions of the estimators in the *non*-null case to draw this conclusion; but see also Section 5.1.

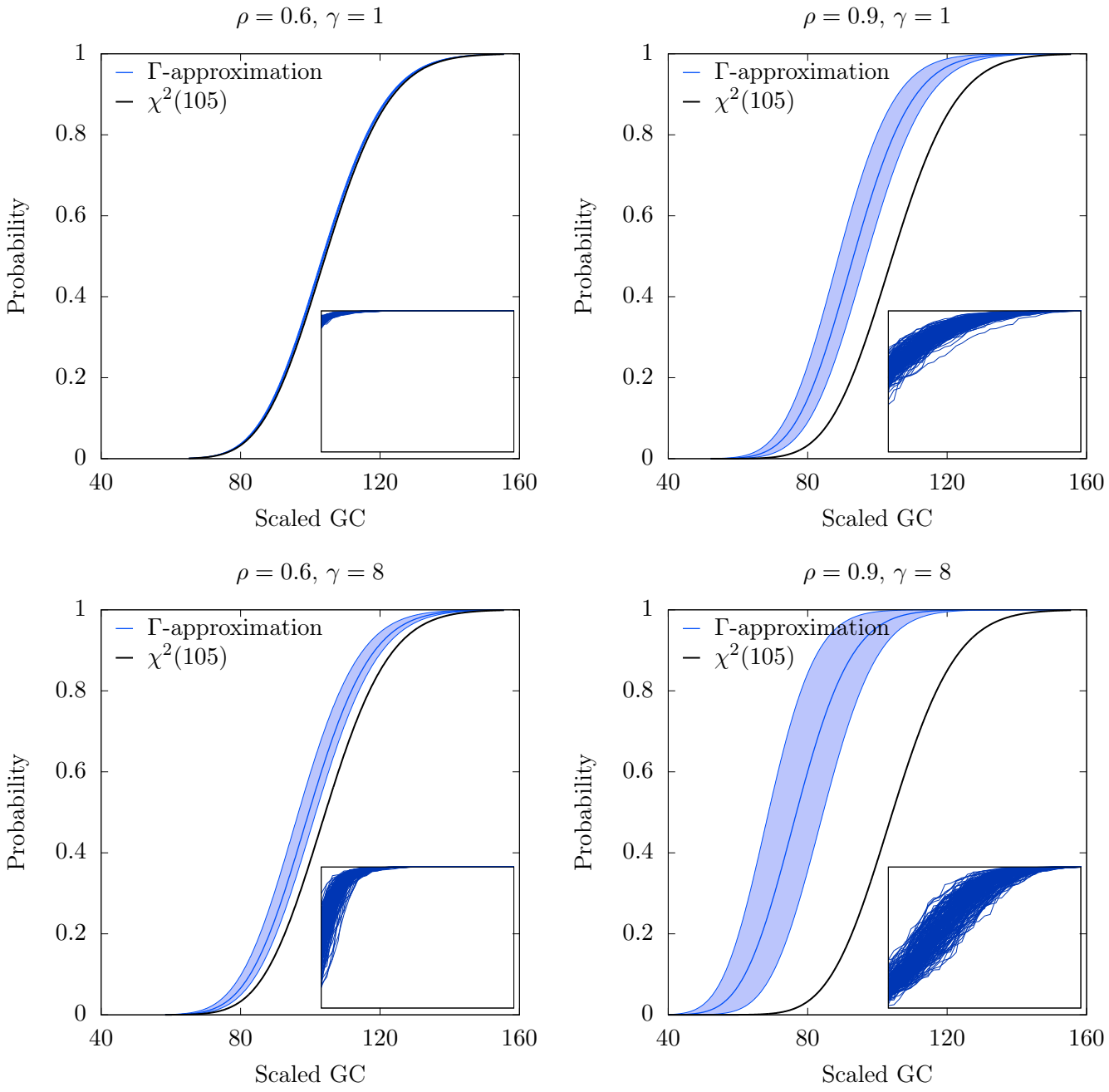

Figure 2: Distribution of Γ-approximation CDFs for a random sample of 200 null VAR models (Appendix D) with $n_x = 3$, $n_y = 5$ and $p = 7$, for a selection of spectral radii ρ and residuals generalised correlation γ. At each scaled GC value, blue lines plot the mean of the Γ-approximation CDFs, while shaded areas bound upper/lower 95% quantiles. Black lines plot the corresponding LR χ^2 CDFs, with $pn_xn_y = 105$ degrees of freedom. Inset figures: the $pn_y = 35$ distinct eigenvalues sorted by size, for each of the 200 generalised χ^2 distributions (x range is $1 - 35$, y range is $0 - 1$).

not vanish, so that $\Psi(\omega)$—and hence, via (28), $S(\omega)$ and consequently $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$—are analytic functions of the phase angle ω [as well as of the $\boldsymbol{\theta} = (A, \Sigma)$]. For a frequency range $\mathcal{F} \subseteq [0, 2\pi]$ with measure $|\mathcal{F}| > 0$, then, $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$ vanishes iff $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ is identically zero; i.e., precisely under the original null hypothesis H_0. Given a frequency range $\mathcal{F}$, we thus apply Proposition 3.1 to the asymptotic distribution of the band-limited spectral Granger causality estimator $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$ under the (appropriate) original null hypothesis (37) $H_0 : A_{k,xy} = 0\ \forall k$.

In the previous section we calculated the covariance $\Omega_0(\boldsymbol{\theta}) = \Sigma_{xx} \otimes \left[\boldsymbol{\Gamma}^{-1}\right]_{yy}$ of null parameters under H_0; it remains to calculate the null Hessian $\mathbf{H}_0(\mathcal{F};\boldsymbol{\theta})$. Since (Lebesgue) integration and partial differentiation are linear operations, it follows that the Hessian for $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$—on the original null space Θ_0—is just

$$\mathbf{H}_0(\mathcal{F};\boldsymbol{\theta}) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} \mathbf{H}_0(\omega;\boldsymbol{\theta})\, d\omega \tag{67}$$

where, for given ω, $\mathbf{H}_0(\omega;\boldsymbol{\theta})$ is the Hessian of $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ on Θ_0 with respect to the original null parameters $A_{k,xy}$, $k = 1 \ldots, p$.

Dropping the "ω" and "$\boldsymbol{\theta}$" arguments for compactness where convenient, on the null space Θ_0 we have $\Phi_{xy} = 0$, and since then Φ is lower block-triangular, we have also $\Psi_{xx} = [\Phi_{xx}]^{-1}$, $\Psi_{yy} = [\Phi_{yy}]^{-1}$, and $\Psi_{xy} = 0$. The CPSD for the process $\boldsymbol{X}_t$ is given by

$$S_{xx} = [\Psi S \Psi^*]_{xx} = \Psi_{xx}\Sigma_{xx}\Psi_{xx}^* + \Psi_{xy}\Sigma_{yx}\Psi_{xx}^* + \Psi_{xx}\Sigma_{xy}\Psi_{xy}^* + \Psi_{xy}\Sigma_{yy}\Psi_{xy}^* \tag{68}$$

On the null space $S_{xx} = \Psi_{xx}\Sigma_{xx}\Psi_{xx}^*$ so that $[S_{xx}]^{-1} = \Phi_{xx}^*[\Sigma_{xx}]^{-1}\Phi_{xx}$,

We define $T(\omega)$ to be the $n_x \times n_x$ (Hermitian) matrix

$$T(\omega) = \Psi_{xy}(\omega)\Sigma_{yy|x}\Psi_{xy}(\omega)^* \tag{69}$$

so that from (32) $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega) = \log|S_{xx}(\omega)| - \log|S_{xx}(\omega) - T(\omega)|$. $T(\omega)$ vanishes on the null space. We may check that (for $p, q, r, s = 1, \ldots, n$, $k = 1, \ldots, p$)

$$\frac{\partial \Psi_{pq}}{\partial A_{k,rs}} = \Psi_{pr}\Psi_{sq}e^{-i\omega k} \tag{70}$$

from which we may calculate (with $i, i', u, u' \in x$, $j, j', v, v' \in y$)

$$\frac{\partial T_{ii'}}{\partial A_{k,uv}} = \sum_{j,j'}[\Sigma_{yy|x}]_{jj'}\left(\Psi_{iu}\Psi_{vj}\bar{\Psi}_{i'j'}e^{-i\omega k} + \bar{\Psi}_{i'u}\bar{\Psi}_{vj'}\Psi_{ij}e^{i\omega k}\right) \tag{71}$$

so that in particular, $\left.\dfrac{\partial T}{\partial \theta_\alpha}\right|_{\boldsymbol{\theta}\in\Theta_0} = 0$ for a null parameter θ_α, and we find [*cf.* (55)]:

$$\left.\frac{\partial^2 f_{\boldsymbol{Y}\to\boldsymbol{X}}}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\boldsymbol{\theta}\in\Theta_0} = \operatorname{trace}\left[[S_{xx}]^{-1}\left.\frac{\partial^2 T}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\boldsymbol{\theta}\in\Theta_0}\right] \tag{72}$$

for null parameters $\theta_\alpha, \theta_\beta$. From (71) and using (70), we may calculate

$$\left.\frac{\partial^2 T_{ii'}}{\partial A_{k,uv}\partial A_{k',u'v'}}\right|_{\boldsymbol{\theta}\in\Theta_0} = \Psi_{iu}\Psi_{u'i'}^*\left[S_{yy|x}\right]_{vv'}e^{-i\omega(k-k')} + \Psi_{iu'}\Psi_{ui'}^*\left[S_{yy|x}\right]_{v'v}e^{i\omega(k-k')} \tag{73}$$

where

$$S_{yy|x} = \Psi_{yy}\Sigma_{yy|x}\Psi_{yy}^* \tag{74}$$

is the CPSD for a VAR(p) model with parameters $(A_{yy}, \Sigma_{yy|x})$ [*cf.* Section 2.4]. From (72) and (73) we find

$$\left.\frac{\partial^2 f_{\boldsymbol{Y}\to\boldsymbol{X}}}{\partial A_{k,uv}\partial A_{k',u'v'}}\right|_{\boldsymbol{\theta}\in\Theta_0} = \left[[\Sigma_{xx}]^{-1}\right]_{uu'}\left\{\left[S_{yy|x}\right]_{vv'}e^{-i\omega(k-k')} + \left[S_{yy|x}\right]_{v'v}e^{i\omega(k-k')}\right\} \tag{75}$$

$$= \left[[\Sigma_{xx}]^{-1}\right]_{uu'}\left[S_{yy|x}e^{-i\omega(k-k')} + \bar{S}_{yy|x}e^{i\omega(k-k')}\right]_{vv'} \quad \text{since } S_{yy|x} \text{ is Hermitian} \tag{76}$$

$$= 2\left[[\Sigma_{xx}]^{-1}\right]_{uu'}\mathfrak{Re}\left\{\left[\boldsymbol{S}_{yy|x}\right]_{kk',vv'}\right\} \tag{77}$$

where for given ω we define the $pn_y \times pn_y$ Hermitian matrix

$$\left[\boldsymbol{S}_{yy|x}(\omega)\right]_{kk',vv'} = [S_{yy|x}(\omega)]_{vv'}e^{-i\omega(k-k')} \tag{78}$$

or

$$\boldsymbol{S}_{yy|x}(\omega) = \boldsymbol{Z}\otimes S_{yy|x}(\omega)\,, \qquad \boldsymbol{Z}_{kk'} = e^{-i\omega(k-k')} \tag{79}$$

We note that $\boldsymbol{S}_{yy|x}(\omega)$ is the CPSD for the "companion VAR(1)" [*cf.* (17)] of the VAR(p) model with parameters $(A_{yy}, \Sigma_{yy|x})$, and as such may be thought of as the spectral counterpart of the autocovariance matrix $\boldsymbol{\Gamma}_{yy|x}$ of Section 3.1. We thus have, for $\omega \in [0, 2\pi]$:

$$\mathbf{H}_0(\omega;\boldsymbol{\theta}) = [\Sigma_{xx}]^{-1}\otimes\mathfrak{Re}\left\{\boldsymbol{S}_{yy|x}(\omega)\right\} \tag{80}$$

so that

$$\mathbf{H}_0(\mathcal{F};\boldsymbol{\theta}) = [\Sigma_{xx}]^{-1}\otimes\mathfrak{Re}\left\{\boldsymbol{S}_{yy|x}(\mathcal{F})\right\} \tag{81}$$

with

$$\boldsymbol{S}_{yy|x}(\mathcal{F}) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}}\boldsymbol{S}_{yy|x}(\omega)\,d\omega \tag{82}$$

We may thus state

Theorem 3.3. The asymptotic distribution of the single-regression band-limited Granger causality estimator over a frequency range $\mathcal{F}\subseteq[0,2\pi]$ under the null hypothesis $H_0 : A_{k,xy} = 0\,\forall k$, is given by

$$N\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta}) \xrightarrow{d} \chi^2\left(\boldsymbol{I}_{xx}\otimes\mathfrak{Re}\left\{\boldsymbol{S}_{yy|x}(\mathcal{F})\right\}, \boldsymbol{I}_{xx}\otimes\left[\boldsymbol{\Gamma}^{-1}\right]_{yy}\right) \quad \text{as } N\to\infty \tag{83}$$

where $\boldsymbol{S}_{yy|x}(\mathcal{F})$ is given by (74), (79) and (82), and $\boldsymbol{\Gamma}$ is as in Theorem 3.1. ■

In particular, for $\mathcal{F} = [0, 2\pi]$ we have

$$\begin{aligned}
[\boldsymbol{S}_{yy|x}([0,2\pi])]_{kk',vv'} &= \frac{1}{2\pi}\int_{\omega=0}^{2\pi}[\boldsymbol{S}_{yy|x}(\omega)]_{kk',vv'}\,d\omega \\
&= \frac{1}{2\pi}\int_{\omega=0}^{2\pi}[S_{yy|x}(\omega)]_{vv'}e^{-i\omega(k-k')}d\omega \\
&= [\Gamma_{yy|x}]_{k'-k,vv'} \qquad \text{by (27)} \\
&= [\boldsymbol{\Gamma}_{yy|x}]_{kk',vv'}
\end{aligned}$$

so that

$$\boldsymbol{S}_{yy|x}([0,2\pi]) = \boldsymbol{\Gamma}_{yy|x} \tag{84}$$

Thus we confirm that the distribution of $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$ is consistent with Theorem 3.1 and (34) for $\mathcal{F} = [0, 2\pi]$. We note that the limiting asymptotic distribution of $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}([\omega-\varepsilon,\omega+\varepsilon];\boldsymbol{\theta})$ as $\varepsilon\to 0$ is obtained by simply replacing $\boldsymbol{S}_{yy|x}(\mathcal{F})$ by $\boldsymbol{S}_{yy|x}(\omega)$ in (83); as mentioned, this should not be confused with the distribution of the point-frequency estimator $\hat{f}_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ under the null hypothesis (115).

As before, the generalised χ^2 distribution in (83) may be described in terms of the eigenvalues λ_i of $\left[\mathbf{\Gamma}^{-1}\right]_{yy}\boldsymbol{S}_{yy|x}(\mathcal{F})$. For $p > 2$, since the null space of the null hypothesis H_0 used to derive (83) has dimension pn_xn_y—whereas for any given angular frequency ω the dimension of the point-frequency null $H_0(\omega)$ (Section 5.2) is $2n_xn_y$—only $2n_y$ of the pn_y eigenvalues of $\left[\mathbf{\Gamma}^{-1}\right]_{yy}\boldsymbol{S}_{yy|x}(\omega)$ will be non-zero. In contrast to the time-domain case, the eigenvalues of $\left[\mathbf{\Gamma}^{-1}\right]_{yy}\boldsymbol{S}_{yy|x}(\mathcal{F})$ will not necessarily be ≤ 1 (*cf.* Conjecture 3.2), but empirically we observe that the maximum eigenvalue shrinks to 1 as the bandwidth $|\mathcal{F}|$ increases to 2π.

3.3 Worked example: the general bivariate VAR(1)

In Appendix C we analyse the bivariate VAR(1)

$$X_t = a_{xx}X_{t-1} + a_{xy}Y_{t-1} + \varepsilon_{xt} \tag{85a}$$
$$Y_t = a_{yx}X_{t-1} + a_{yy}Y_{t-1} + \varepsilon_{yt} \tag{85b}$$

with residuals covariance matrix $\mathbf{E}\left[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\mathsf{T}\right] = \begin{bmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{yx} & \sigma_{yy}\end{bmatrix}$, so that $\boldsymbol{\theta} = (a_{xx}, a_{xy}, a_{yx}, a_{yy}, \sigma_{xx}, \sigma_{xy}, \sigma_{yy})$. We calculate that the (population) Granger causality from $Y \to X$ is given by

$$F_{Y\to X}(\boldsymbol{\theta}) = \log\frac{v(\boldsymbol{\theta})}{\sigma_{xx}} \tag{86}$$

where the reduced residuals covariance matrix—in this case the scalar $V(\boldsymbol{\theta}) = v(\boldsymbol{\theta})$—is given by

$$v(\boldsymbol{\theta}) = \frac{1}{2}\left(P + \sqrt{P^2 - Q^2}\right) \tag{87}$$

with

$$P = \sigma_{xx}(1 + a_{yy}^2) - 2\sigma_{xy}a_{xy}a_{yy} + \sigma_{yy}a_{xy}^2\,, \qquad Q = 2(\sigma_{xx}a_{yy} - \sigma_{xy}a_{xy}) \tag{88}$$

We apply Theorem 3.1 to calculate the asymptotic distribution of the SR estimator $\hat{F}^{\mathsf{SR}}_{Y\to X}(\boldsymbol{\theta})$ on the null space $a_{xy} = 0$. Noting that for model order $p = 1$, $\mathbf{\Gamma} = \Gamma_0$, and setting $\Gamma_0^{-1} = [\omega_{ij}]$, we have $[\mathbf{\Gamma}^{-1}]_{yy} = [\Gamma_0^{-1}]_{yy} = \omega_{yy}$ [Appendix C, eq. (154)]. The DLYAP equation (61b) for $\Gamma_{yy|x}$ is just

$$\Gamma_{yy|x} - a_{yy}^2\Gamma_{yy|x} = \sigma_{yy|x} \tag{89}$$

with

$$\sigma_{yy|x} = \sigma_{yy} - \frac{\sigma_{xy}^2}{\sigma_{xx}} = \sigma_{yy}\left(1 - \kappa^2\right) \tag{90}$$

where $\kappa = \dfrac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}$ is the residuals correlation, so that

$$\Gamma_{yy|x} = \left(1 - \kappa^2\right)\frac{\sigma_{yy}}{1 - a_{yy}^2} \tag{91}$$

and the single eigenvalue of $[\mathbf{\Gamma}^{-1}]_{yy}\mathbf{\Gamma}_{yy|x}$ is just

$$\lambda = \left(1 - \kappa^2\right)\frac{\sigma_{yy}\omega_{yy}}{1 - a_{yy}^2} \tag{92}$$

By Theorem 3.1 the asymptotic distribution of the single-regression estimator is thus a scaled $\chi^2(1)$:

$$N\hat{F}^{\mathsf{SR}}_{Y\to X}(\boldsymbol{\theta}) \xrightarrow{d} \lambda\cdot\chi^2(1) = \Gamma\left(\frac{1}{2}, 2\lambda\right) \tag{93}$$

(the Γ-approximation in this case is exact).

In Appendix C we calculate the spectral Granger causality from $Y \to X$ at $\omega \in [0, 2\pi]$ as

$$f_{Y\to X}(\omega;\boldsymbol{\theta}) = \log \frac{P - Q\cos\omega}{P - Q\cos\omega - a_{xy}^2\sigma_{yy|x}} \tag{94}$$

We find then that

$$S_{yy|x}(\omega) = \sigma_{yy|x}|\psi_{yy}(\omega)|^2 = \sigma_{yy}\left(1-\kappa^2\right)\frac{|1-a_{xx}z|^2}{|\Delta(\omega)|^2} \tag{95}$$

with $\Delta(\omega) = |\Phi(\omega)|$. On the null space, $\Delta(\omega) = (1-a_{xx}z)(1-a_{yy}z)$, so that

$$S_{yy|x}(\omega) = \left(1-\kappa^2\right)\frac{\sigma_{yy}}{|1-a_{yy}z|^2} = \left(1-\kappa^2\right)\frac{\sigma_{yy}}{1-2a_{yy}\cos\omega + a_{yy}^2} \tag{96}$$

In this case, since the model order is $p = 1$, the null hypothesis (115) coincides with the null hypothesis (37) (i.e., $a_{xy} = 0$), so that from Theorem 3.3 we have

$$N\hat{f}_{Y\to X}(\omega;\boldsymbol{\theta}) \xrightarrow{d} \lambda(\omega)\cdot\chi^2(1) \tag{97}$$

where

$$\lambda(\omega) = \left(1-\kappa^2\right)\frac{\sigma_{yy}\omega_{yy}}{|1-a_{yy}z|^2} = \left(1-\kappa^2\right)\frac{\sigma_{yy}\omega_{yy}}{1-2a_{yy}\cos\omega + a_{yy}^2} \tag{98}$$

The asymptotic distribution for the band-limited estimator may then be calculated as per (82) by integrating (96) across the appropriate frequency range[10].

4 Statistical inference with the single-regression estimators

In this section we show how our principal results, Theorems 3.1 and 3.3, can be usefully applied to inference on Granger causality $F_{\boldsymbol{X}\to\boldsymbol{Y}}$. The key difficulty in this endeavour is the fact that the asymptotic distribution of the single-regression test statistics depends explicitly on the true null parameter, i.e. on its location in the subset of the parameter space associated with the null hypothesis.

Assuming the VAR(p) class of models, and given time-series data $\boldsymbol{u} = \{u_1, u_2, \ldots, u_N\}$, we can estimate a ML parameter $\hat{\boldsymbol{\theta}}(\boldsymbol{u})$; but we cannot simply replace the VAR(p) true null parameter for the asymptotic distribution by the ML estimate, since $\hat{\boldsymbol{\theta}}(\boldsymbol{u})$ cannot be assumed to lie in the null space Θ_0. Our solution is to project the ML parameter estimate onto the null space by a continuous projection $\Pi : \Theta \to \Theta_0$ (so that in particular $\Pi\cdot\boldsymbol{\theta} = \boldsymbol{\theta}$ if $\boldsymbol{\theta} \in \Theta_0$). Here we might choose the obvious projection $\Pi\cdot\boldsymbol{\theta} = (0, \ldots, 0, \theta_{d+1}, \ldots, \theta_r)$. This yields a Neyman-Pearson test for the null hypothesis of zero Granger causality from $\boldsymbol{Y} \to \boldsymbol{X}$, which we term a "Projection Test". Below we show that the Projection Test is "asymptotically valid", in the sense that the Type I error rate (incidence of false rejections of H_0) may be limited to a prespecified level α, and investigate statistical power of the SR estimators in terms of the Type II error rate (incidence of failures to correctly reject H_0) at a given level.

For brevity, in what follows $F(\cdots)$ will refer to either the time-domain population SR statistic $F^{\mathsf{SR}}_{\boldsymbol{X}\to\boldsymbol{Y}}(\cdots)$ (39) or the band-limited population statistic $f_{\boldsymbol{X}\to\boldsymbol{Y}}(\mathcal{F};\cdots)$ (33) for some given frequency range $\mathcal{F}$.

4.1 Type I error rate

Given a null parameter $\boldsymbol{\theta} \in \Theta_0$ under H_0 (37) for a VAR(p) model, let $\Phi_{\boldsymbol{\theta}}(\cdots)$ be the CDF of the corresponding asymptotic null Granger causality estimator; i.e., the generalised χ^2 of (60) (time domain) or (83) (band-limited). Given a data sample $\boldsymbol{u}$ of size N, the order-p Projection Test procedure at level α is as follows:

[10] We may calculate that $\displaystyle\int \frac{d\omega}{1-2a\cos\omega+a^2} = \frac{2}{1-a^2}\tan^{-1}\left(\frac{1+a}{1-a}\tan\frac{\omega}{2}\right)$.

1. Calculate the ML VAR(p) parameter estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{u})$.
2. Calculate the GC estimate $F\big(\hat{\boldsymbol{\theta}}(\boldsymbol{u})\big)$.
3. Reject the null hypothesis if

$$\Phi_{\Pi\cdot\hat{\boldsymbol{\theta}}(\boldsymbol{u})}\Big(NF\big(\hat{\boldsymbol{\theta}}(\boldsymbol{u})\big)\Big) > 1-\alpha \tag{99}$$

We now investigate the asymptotic Type I error rate for this test. Let $\hat{\boldsymbol{\theta}}$ be the ML parameter estimator for a model with true parameter $\boldsymbol{\theta} \in \Theta_0$; recall that we consider $\hat{\boldsymbol{\theta}}$ as a random variable parametrised by $\boldsymbol{\theta}$, with implicit dependence on sample size N, and we note again that $\hat{\boldsymbol{\theta}}$ may take values outside of Θ_0. The GC estimator is $F\big(\hat{\boldsymbol{\theta}}\big)$ (again, a sample size-dependent random variable parametrised by $\boldsymbol{\theta}$). Given $\boldsymbol{\theta} \in \Theta_0$, the probability of a Type I error for the Projection Test is

$$P_I(\boldsymbol{\theta};\alpha) = \mathbf{P}\Big[\Phi_{\Pi\cdot\hat{\boldsymbol{\theta}}}\Big(NF\big(\hat{\boldsymbol{\theta}}\big)\Big) > 1-\alpha\Big] \tag{100}$$

We wish to show that $\lim_{N\to\infty} P_I(\boldsymbol{\theta};\alpha) = \alpha$.

Lemma 4.1. Suppose given a sequence of pairs of real-valued random variables (X_n, Y_n) such that

$$X_n \xrightarrow{d} X \tag{101}$$

$$Y_n \xrightarrow{p} c \tag{102}$$

where c is a constant. Then

$$\mathbf{P}[X_n \le Y_n] \longrightarrow \mathbf{P}[X \le c] \tag{103}$$

Proof. By Slutsky's Lemma (Lehmann and Romano, 2005), we have $X_n - Y_n \xrightarrow{d} X - c$. Thus $\forall\,\varepsilon > 0$, $\exists\, n_0 \in \mathbb{N}$ such that $\forall\, n \ge n_0$ we have $\big|\mathbf{P}[X_n - Y_n \le 0] - \mathbf{P}[X - c \le 0]\,\big| < \varepsilon$, or, equivalently $\big|\mathbf{P}[X_n \le Y_n] - \mathbf{P}[X \le c]\,\big| < \varepsilon$, which establishes (103). □

Now (100) is equivalent to

$$P_I(\boldsymbol{\theta};\alpha) = 1 - \mathbf{P}\Big[NF\big(\hat{\boldsymbol{\theta}}\big) \le \Phi^{-1}_{\Pi\cdot\hat{\boldsymbol{\theta}}}(1-\alpha)\Big] \tag{104}$$

Since $\hat{\boldsymbol{\theta}}$ is a consistent estimator for $\boldsymbol{\theta}$ and Π is continuous with $\Pi\cdot\boldsymbol{\theta} = \boldsymbol{\theta}$, by the Continuous Mapping Theorem (CMT) we have $\Pi\cdot\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$. It is not hard to verify that the inverse CDF $\Phi^{-1}_{\boldsymbol{\theta}}(\cdots)$ evaluated at $1-\alpha$ is continuous in the $\boldsymbol{\theta}$ argument, so that again by the CMT we have $\Phi^{-1}_{\Pi\cdot\hat{\boldsymbol{\theta}}}(1-\alpha) \xrightarrow{p} \Phi^{-1}_{\boldsymbol{\theta}}(1-\alpha)$. By Theorem 3.1, $NF\big(\hat{\boldsymbol{\theta}}\big) \xrightarrow{d} Q_{\boldsymbol{\theta}}$, where $Q_{\boldsymbol{\theta}}$ is a (generalised χ^2) random variable with CDF $\Phi_{\boldsymbol{\theta}}$. Applying Lemma 4.1 to the pair-sequence $\Big(NF\big(\hat{\boldsymbol{\theta}}\big), \Phi^{-1}_{\Pi\cdot\hat{\boldsymbol{\theta}}}(1-\alpha)\Big)$ we have

$$P_I(\boldsymbol{\theta};\alpha) \longrightarrow 1 - \mathbf{P}\big[Q_{\boldsymbol{\theta}} \le \Phi^{-1}_{\boldsymbol{\theta}}(1-\alpha)\big] = 1 - \Phi_{\boldsymbol{\theta}}\big(\Phi^{-1}_{\boldsymbol{\theta}}(1-\alpha)\big) = \alpha \tag{105}$$

as required. While the rate of convergence of $P_I(\boldsymbol{\theta};\alpha)$ to α can be expected to depend on the true null parameter $\boldsymbol{\theta}$, empirically we find that the effect is small[11].

To gain insight into comparative Type I error rates and their parameter dependencies, we performed the following experiment: at each data length N, we drew 1,000 model samples $\boldsymbol{\theta}$ from a distribution $\tilde{\boldsymbol{\theta}}$ over the subspace $\Theta(\rho,\gamma) \subset \Theta$ of null VAR models of given size, for a given spectral radius ρ and residuals generalised correlation γ, according to the scheme described in Appendix D. For each model $\boldsymbol{\theta}$,

[11] The extent to which the choice of projection Π may affect the rate of convergence is unclear. It is plausible that orthogonal projection with respect to the Fisher metric (Amari, 2016) evaluated at the null parameter may lead to faster convergence, but we have not tested this hypothesis.

we generate $n = 1,000$ data sequences. For each data sequence we calculate LR and SR estimates of the Granger causality, and test the estimates against the appropriate χ^2 (resp. generalised χ^2) asymptotic null distribution at significance level $\alpha = 0.05$. For a given VAR model $\boldsymbol{\theta}$, let $p(\boldsymbol{\theta})$ be the population Type I error rate associated with the given data length, estimator, inference method and significance level. Each of the n statistical tests on a given model is a Bernoulli trial, yielding a binomially-distributed estimator $\hat{p}(\boldsymbol{\theta})$ for $p(\boldsymbol{\theta})$; i.e., $n\hat{p}(\boldsymbol{\theta}) \sim B\big(n, p(\boldsymbol{\theta})\big)$:

$$\mathbf{P}[n\hat{p}(\boldsymbol{\theta}) = k] = \binom{n}{k} p(\boldsymbol{\theta})^k [1 - p(\boldsymbol{\theta})]^{n-k}, \qquad k = 0, \ldots, n \tag{106}$$

The $\hat{p}(\boldsymbol{\theta})$ estimates are collated over model samples from $\tilde{\boldsymbol{\theta}}$, to yields a distribution $\hat{p}$ for each data length N; explicitly,

$$\mathbf{P}\Big[n\hat{p} = k \,|\, \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}\Big] = \mathbf{P}[n\hat{p}(\boldsymbol{\theta}) = k], \qquad k = 0, \ldots, n \tag{107}$$

Results for $n_x = 3$, $n_y = 5$, $p = 7$, $\gamma = 1$, $N = 2^8 - 2^{14}$ and a selection of spectral radii are presented in Figure 3. We see that asymptotic convergence for the SR estimator Type I error rate is slightly better on average than for the LR estimator, especially for smaller sample lengths and spectral radius $\rho(A)$ (20) close to 1.

Note that there are two sources of variation in the experiment: estimation error of the per-model error rate, represented by the estimators $\hat{p}(\boldsymbol{\theta})$, and variation across the distribution $\tilde{\boldsymbol{\theta}}$ of models. From the Law of Total Variance, we may calculate that

$$\mathrm{var}[\hat{p}] = \mathrm{var}_{\tilde{\boldsymbol{\theta}}}\Big[p(\tilde{\boldsymbol{\theta}})\Big] + \frac{1}{n}\mathbf{E}_{\tilde{\boldsymbol{\theta}}}\Big[p(\tilde{\boldsymbol{\theta}})[1 - p(\tilde{\boldsymbol{\theta}})]\Big] \tag{108}$$

We note that the second term is $\leq \frac{1}{4n}$, so that for large enough n the $\mathrm{var}_{\tilde{\boldsymbol{\theta}}}\Big[p(\tilde{\boldsymbol{\theta}})\Big]$ term—the variance of Type II error rate with respect to model variation—dominates; i.e., for large enough n the error bars are essentially due to the effect of model variation on the GC estimates. In Figure 3, where the contribution to the error bars from the per-model rate estimators $\hat{p}(\boldsymbol{\theta})$ is small (of the order[12] $\frac{1}{\sqrt{n}} \approx 0.032$), the visible dispersion is thus mostly due to model variation $\tilde{\boldsymbol{\theta}}$.

4.2 Type II error rate (statistical power)

As regards statistical power, the Type II error rate given a *non*-null parameter $\boldsymbol{\theta} \in \Theta$, is given by

$$P_{II}(\boldsymbol{\theta}; \alpha) = \mathbf{P}\Big[\Phi_{\Pi \cdot \hat{\boldsymbol{\theta}}}\Big(NF(\hat{\boldsymbol{\theta}})\Big) \leq 1 - \alpha\Big] \tag{109}$$

For the likelihood-ratio statistic we have the classical result due to Wald (1943), which yields that the scaled estimator is $\approx$ non-central $\chi^2(pn_xn_y; NF)$ in the large-sample limit, where F is the population GC. The approximation only holds with reasonable accuracy for "small" values of F. In the single-regression case we have no equivalent result (but see discussion in Section 5). Clearly, $P_{II}(\boldsymbol{\theta}; \alpha)$ will depend strongly on the population GC value associated with specific parameters $\boldsymbol{\theta}$, but, as for the null case, will still vary within the subspace of parameters which yield a given population statistic; that is, for given $F > 0$, $P_{II}(\boldsymbol{\theta}; \alpha)$ will vary over the set $\{\boldsymbol{\theta} : F(\boldsymbol{\theta}) = F\}$. We can at least say that, roughly

$$P_{II}(\boldsymbol{\theta}; \alpha) \approx 0 \quad \text{for} \quad N \gg \frac{\Phi^{-1}_{\boldsymbol{\theta}}(1 - \alpha)}{F(\boldsymbol{\theta})} \tag{110}$$

$$P_{II}(\boldsymbol{\theta}; \alpha) \approx 1 \quad \text{for} \quad N \ll \frac{\Phi^{-1}_{\boldsymbol{\theta}}(1 - \alpha)}{F(\boldsymbol{\theta})} \tag{111}$$

[12] 95% quantiles correspond approximately to two standard deviations.

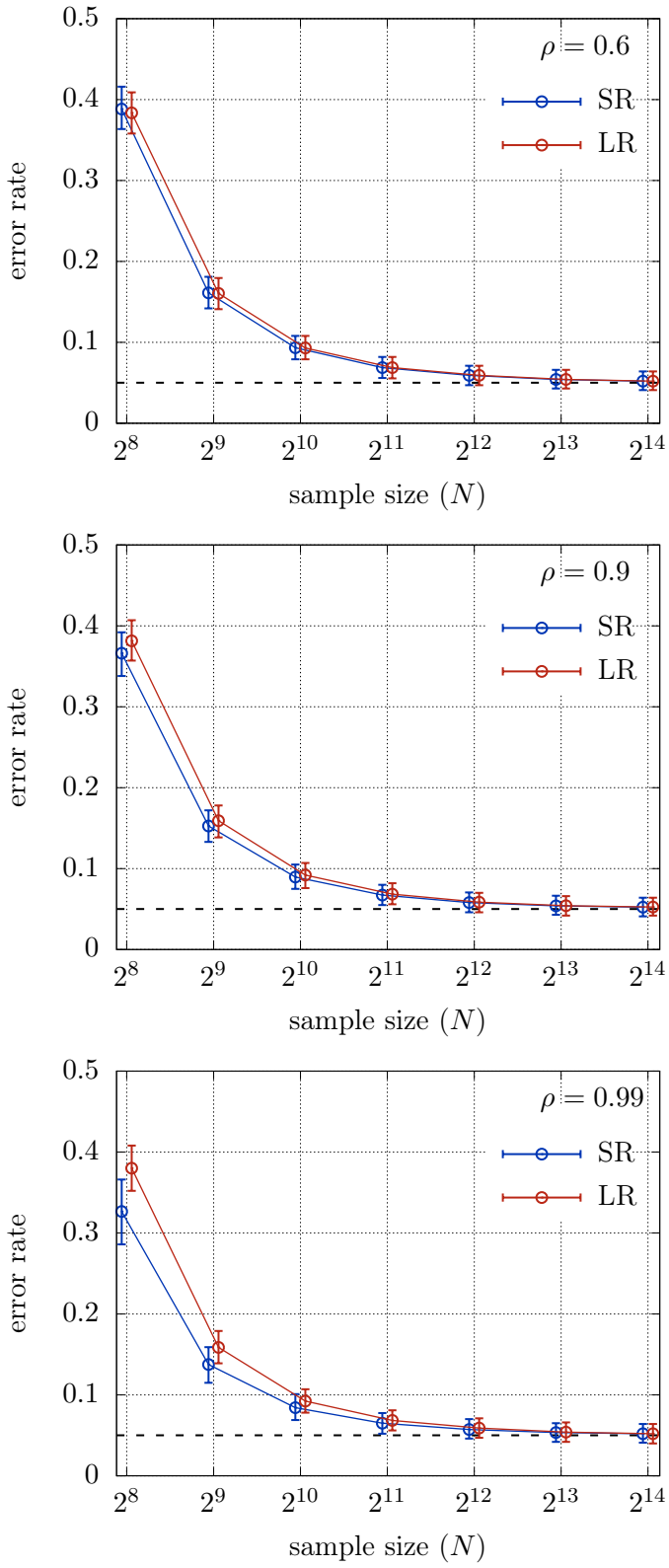

Figure 3: Type I error rates at significance level $\alpha = 0.05$ for the SR vs. LR estimator for null VAR models with $n_x = 3$, $n_y = 5$, $p = 7$, $\gamma = 1$ and spectral radius ρ as indicated, plotted against data length N (note log scale). Error bars represent 95% upper/lower quantiles; the dashed horizontal line indicates the significance level. See text (Section 4.1) for simulation details.

To compare statistical power (i.e., 1−Type II error rate) between the SR and LR estimators, and to examine their dependencies on model parameters, we performed simulations on the bivariate VAR(1) of Section 3.3. Results are displayed in Figure 4. We see (left-column figures) that for the SR estimator, statistical power is roughly symmetrical around $a_{xx} = a_{yy}$ (we note that in this case, spectral radius is $\max\{|a_{xx}|, |a_{yy}|\}$, so spectral radius "swaps" between a_{xx} and a_{yy} along this line). Statistical power is highest when a_{xx} and a_{yy} are roughly equal in magnitude and opposite in sign. Parameter dependency is stronger for larger residuals correlation κ. The power difference (right-column figures) between the estimators is small when residuals correlation is small; for larger residuals correlation, there is again rough (anti-)symmetry in the power difference around $a_{xx} = a_{yy}$, but the dependencies are quite complex.

To gain some insight into comparative statistical power on more general VAR models, we also repeated the experiment described in Section 4.1, but this time generating non-null models with population GC $F = 0.007$ (see Appendix D), and testing for Type II errors. Results are presented in Figure 5. The overall conclusion is that the SR and LR tests have similar statistical power, except when spectral radius is very close to 1, where the LR test has somewhat higher power. Again, visible dispersion is mostly due to model variation $\tilde{\boldsymbol{\theta}}$.

Figure 5 raises a puzzling point: we might suspect that the distribution of $\hat{p}$ for the LR statistic has *small* variance due to $\tilde{\boldsymbol{\theta}}$, since—at least in the null case—the asymptotic distribution depends only on the problem size (degrees of freedom). This is not what our simulations show, however (see in particular $\rho = 0.99$ and $2^{10} \leq N \leq 2^{12}$). But note that in the non-null case this may not apply; if we are in the "Wald regime", where F is "small", then similar should apply, since the Wald non-central χ^2 non-null asymptotic distribution again depends only on degrees of freedom (and the actual GC value F). We can only conclude that for the particular parameters of the experiment, F lies beyond the ambit of Wald's Theorem, so that somewhat counter-intuitively—that is, despite the known dependence of the SR sampling distribution in the null case on VAR parameters—variance of Type II error rate is generally *smaller* than for the LR case.

5 Discussion

In this study we have derived, for the first time, asymptotic generalised χ^2 sampling distributions for the (unconditional) single-regression Granger causality in the time domain, and for the band-limited single-regression estimator in the frequency domain. We conclude with some discussion on implications, limitations and future extensions of this work.

5.1 Unknown and infinite VAR model order

So far we have assumed that the model order of the underlying VAR model is both finite and known. However, these restrictions will generally not be met in practice. The question, then, is how statistical inference is affected when the true model order is unknown, infinite or both. Apart from some general remarks (see below), we consider here only the case of finite, but unknown VAR model order p.

If the model order is unknown, statistical inference becomes a two-stage process: first we obtain a parsimonious model order estimate $\hat{p}$ using a standard selection procedure (McQuarrie and Tsai, 1998). We then compute a test-statistic $F^{(\hat{p})}$—the log-likelihood ratio or single-regression estimate—using the selected model order; that is, maximum-likelihood VAR($\hat{p}$) parameter estimates are computed for the full (and in the case of the LR estimator, reduced) models.

The central question is how the model order selection step should be performed so that the two-step testing procedure as a whole is (1) asymptotically valid, and (2) is as statistically powerful as possible. We consider first the implications of selecting a fixed model order $q \neq p$ for inference on a VAR(p) process. If $q < p$, then the asymptotic statements underlying the likelihood-ratio test (i.e. Wilks' theorem) and projection test (Theorem 3.1 above) no longer hold; thus, for instance, in the case of the LR statistic, $NF^{(q)} \xrightarrow{d} \chi^2(qn_x n_y)$ fails under the null hypothesis (37). On the other hand, if $q > p$ then Wilks' theorem

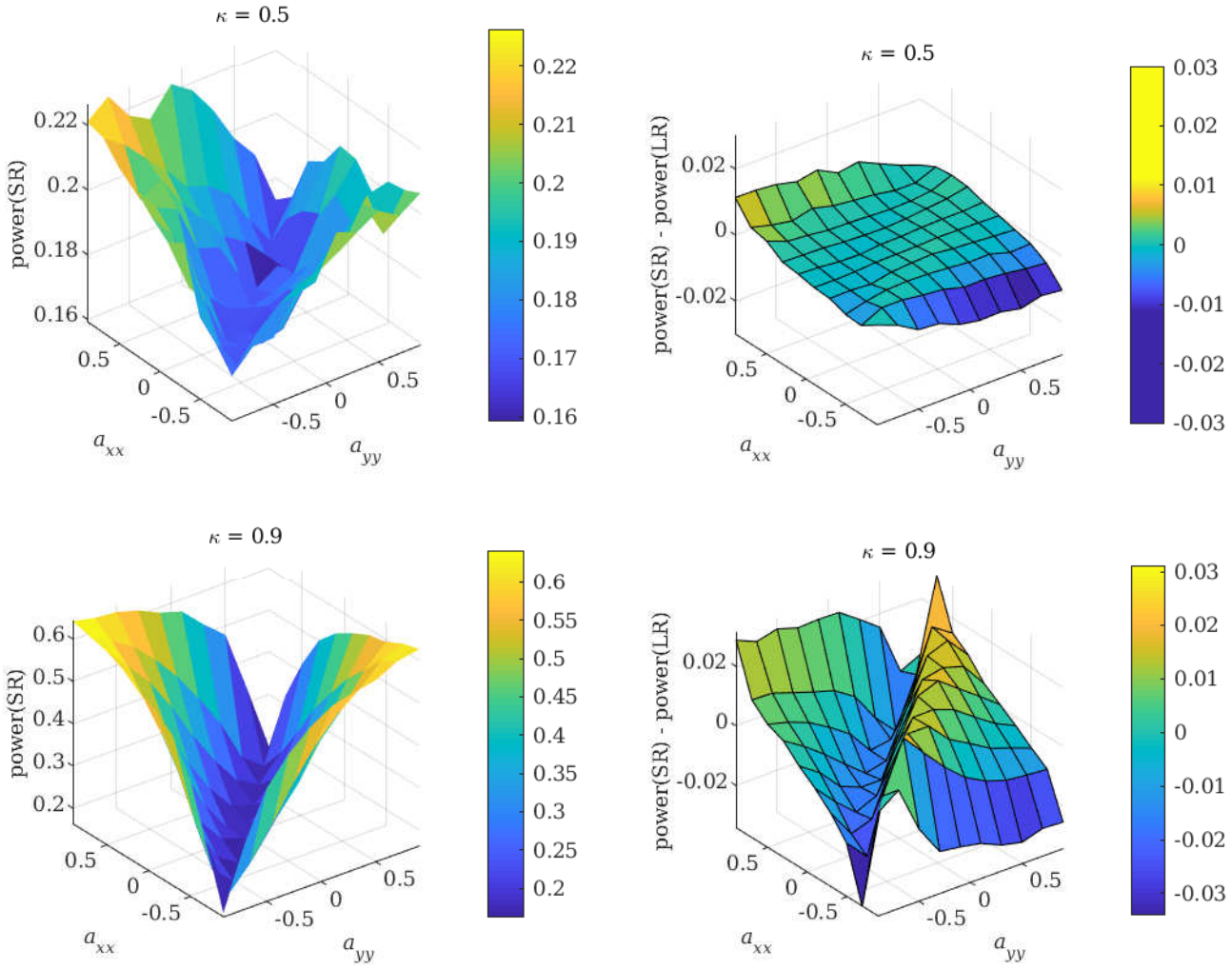

Figure 4: Statistical power (1 − Type II error rate) for the bivariate VAR(1) (85, Section 3.3), plotted against null-space parameters a_{xx}, a_{yy} for residuals correlation $\kappa = 0.5$ and $\kappa = 0.9$. Other parameters were $N = 10^4$, $a_{yx} = 0$, $F = 10^{-4}$. Results were based on 10^4 simulated VAR(1) sequences for each set of null-space parameters. Left column: statistical power for the SR estimator. Right column: difference in statistical power (SR power − LR power).

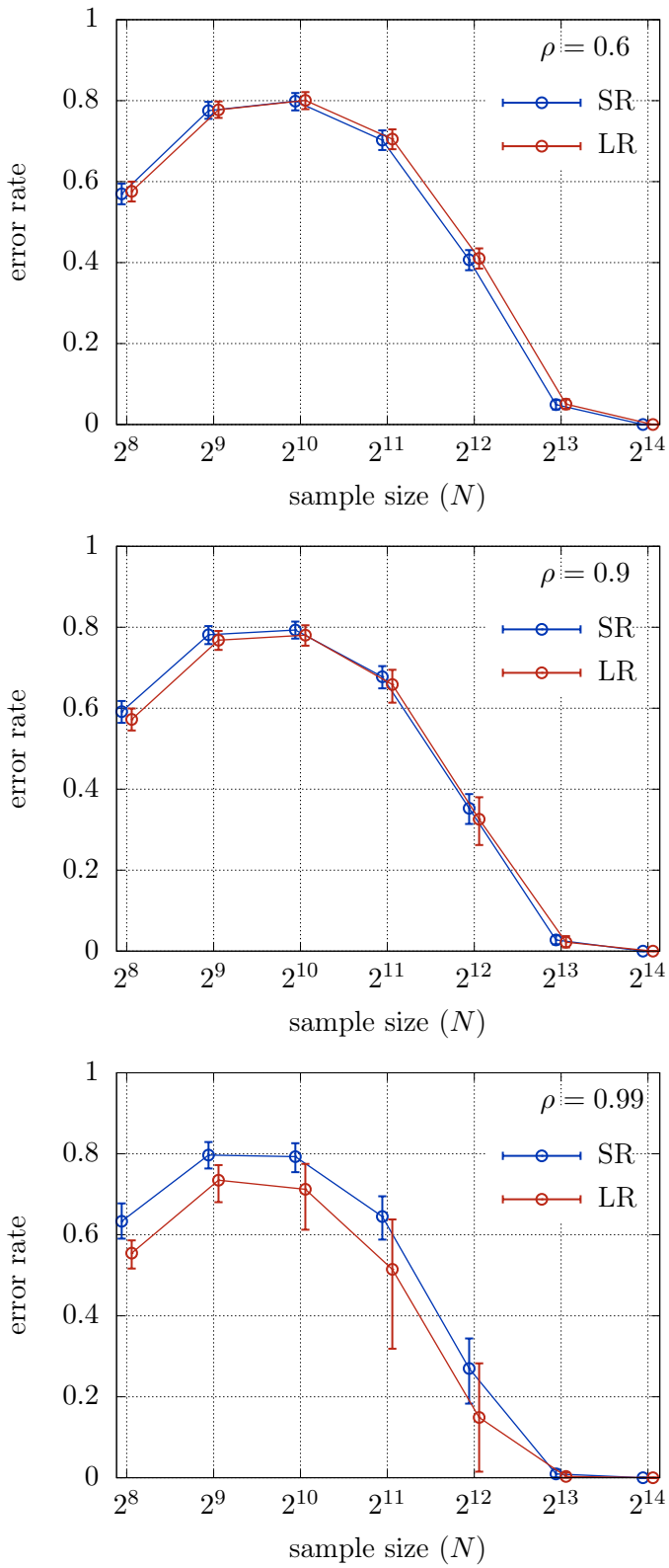

Figure 5: Type II error rates (1 − statistical power) at $\alpha = 0.05$ for the SR vs. LR estimator for causal VAR models with $n_x = 3$, $n_y = 5$, $p = 7$, $\gamma = 1$, population GC $F = 0.007$, and spectral radius ρ as indicated; other details as in Figure 3.

does apply, because (*cf.* Section 2.4.1) one can always subsume a VAR model by a higher-order model. Thus, under the null hypothesis, the asymptotic statement $NF^{(q)} \xrightarrow{d} \chi^2(qn_xn_y)$ is correct for fixed $q > p$, even though the model is over-specified. However, simulation results suggest two problems associated with using too-large a model order: firstly, the rate of convergence of the test statistic decreases with q—potentially leading to inflated Type I errors in small samples—and secondly, statistical power is degraded.

These considerations imply that *for the purposes of statistical inference*, model order $\hat{p}$ should be selected for the *full*, rather than reduced process. Because the reduced process will in general be infinite order, the reduced model order estimate will diverge to infinity as sample size increases, leading to suboptimal inference as described above. For estimation of Granger causality *as an effect size*, however, *efficiency* of the estimate takes precedence. For the SR estimator there is again no reason not to select for model order based on the full process. For the LR estimator, although for fixed sample size N choice of full vs. reduced model order selection runs into the bias-variance trade-off identified by Stokes and Purdon (2017) (Section 2.4.1), empirical investigation reveals that, if model order $\hat{p}$ is selected for the reduced process by a standard scheme, both bias and variance tend to zero as $N \to \infty$. This is explained by the observation that $\hat{p}$ grows comparatively slowly with increasing N. Indeed, if the population Granger causality $F > 0$ is small enough that the large-sample non-central χ^2 approximation of Wald (1943) is viable for the LR estimator, both bias and variance are $\boldsymbol{O}\left(\hat{p}N^{-1}\right)$, while in general $\hat{p}$ will grow more slowly than $\boldsymbol{O}(N)$ (Hannan and Deistler, 2012, Chap. 5). Further research is required here, though ultimately the issue is of little import since under the effect-size scenario there seems scant reason *not* to prefer the SR estimator on grounds of efficiency; *cf.* the closing remarks in Section 3.1.

Given that model order is selected based on the full process, there is still a choice to be made regarding which of the many possible selection criteria should be deployed. In this regard we note that if the model order selection criterion utilised is *consistent*[13] then the probability of choosing the correct model order converges to 1 as $N \to \infty$. In this case the asymptotic results underlying the likelihood-ratio test and single regression projection test remain valid even if the model order has to be estimated:

$$\begin{aligned}
\lim_{N\to\infty} \mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon\right] &= \lim_{N\to\infty} \left(\mathbf{P}[\hat{p} = p]\,\mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon \;\middle|\; \hat{p} = p\right] + \mathbf{P}[\hat{p} \neq p]\,\mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon \;\middle|\; \hat{p} \neq p\right]\right) \\
&= \lim_{N\to\infty} \mathbf{P}[\hat{p} = p]\,\mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon \;\middle|\; \hat{p} = p\right] + \lim_{N\to\infty} \mathbf{P}[\hat{p} \neq p]\,\mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon \;\middle|\; \hat{p} \neq p\right] \\
&= \lim_{N\to\infty} \mathbf{P}[\hat{p} = p]\,\mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon \;\middle|\; \hat{p} = p\right] + 0 \\
&= \lim_{N\to\infty} \mathbf{P}\left[F^{(\hat{p})} \leq \varepsilon \;\middle|\; \hat{p} = p\right]
\end{aligned}$$

We have shown (Theorem 3.1 and Section 4.1) that for the Projection Test the final limit is the CDF of the corresponding generalised χ^2 distribution evaluated at ε, while for the LR test, by Wilks' theorem it is the CDF of the corresponding χ^2 distribution evaluated at ε. Hence, the convergence statements justifying these tests still hold, and the corresponding two-step procedures including the model selection step are asymptotically valid.

As regards the infinite-order VAR case, establishing an asymptotically-valid scheme would seem to be more difficult, and merits further research. Preliminary experiments indicate that, at least for VARMA (equivalently state-space) processes, the (autoregressive) SR estimator with consistent model order selection yields asymptotically valid inference, in the sense that the Type I error rate converges to a specified significance level α (*cf.* Section 4.1), and is also more efficient than the LR estimator. This seems reasonable, given the possibility (see Section 5.2 below) of extending the analysis in this study to the state-space GC estimator (Barnett and Seth, 2015; Solo, 2016).

[13] Note that the popular Akaike Information Criterion (AIC) is *not* consistent, whereas, e.g., Schwartz' Bayesian Information Criterion (BIC) and Hannan amd Quinn's Information Criterion (HQIC; Hannan and Quinn, 1979) are consistent.

5.2 Extensions and future research directions

We make the observation here that *any* estimator of the form $f(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the ML parameter estimator, will converge in distribution to a generalised χ^2 distribution under the associated null hypothesis $\boldsymbol{\theta} \in \Theta_0$ if the statistic $f(\boldsymbol{\theta})$ satisfies the prerequisites for Proposition 3.1. This covers a range of extensions to our results, which we list below. They vary in tractability according to the difficulty of explicit calculation of the Fisher information matrix $\Omega_0(\boldsymbol{\theta})$ and Hessian $\mathbf{H}_0(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0$.

- **The conditional case**

  Extending the time-domain Theorem 3.1 to the conditional case (Geweke, 1984) is reasonably straightforward. Given a partitioning $\boldsymbol{U}_t = [\boldsymbol{X}_t^\mathsf{T} \ \boldsymbol{Y}_t^\mathsf{T} \ \boldsymbol{Z}_t^\mathsf{T}]^\mathsf{T}$ of the variables, the time-domain conditional population Granger causality statistic is given by

$$F_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{\theta}) = \log \frac{\left|\Sigma^\mathsf{R}_{xx}\right|}{\left|\Sigma_{xx}\right|} \tag{112}$$

  where now the reduced regression is

$$\boldsymbol{X}_t = \sum_{k=1}^{\infty} A^\mathsf{R}_{k,xx}\boldsymbol{X}_{t-k} + A^\mathsf{R}_{k,xz}\boldsymbol{Z}_{t-k} + \boldsymbol{\varepsilon}^\mathsf{R}_{x,t} \tag{113}$$

$$\boldsymbol{Z}_t = \sum_{k=1}^{\infty} A^\mathsf{R}_{k,zx}\boldsymbol{X}_{t-k} + A^\mathsf{R}_{k,zz}\boldsymbol{Z}_{t-k} + \boldsymbol{\varepsilon}^\mathsf{R}_{z,t} \tag{114}$$

  with residuals covariance matrix $\Sigma^\mathsf{R} = \mathbf{E}\left[\boldsymbol{\varepsilon}^\mathsf{R}_t \boldsymbol{\varepsilon}^{\mathsf{RT}}_t\right] = \begin{bmatrix} \Sigma^\mathsf{R}_{xx} & \Sigma^\mathsf{R}_{xz} \\ \Sigma^\mathsf{R}_{zx} & \Sigma^\mathsf{R}_{zz} \end{bmatrix}$. Again, Σ^R is a deterministic (albeit more complicated) function $V(\boldsymbol{\theta})$ of the VAR parameters, which may again be expressed in terms of a DARE (Barnett and Seth, 2015). Although more complex, derivation of the appropriate Hessian proceeds along the same lines as in Section 3.1.

  Extension to the conditional case in the frequency domain (band-limited estimator) is substantially more challenging, due to the complexity of the statistic; see e.g., Barnett and Seth (2015, Sec. II). Note that while the unconditional spectral statistic only references the full model parameters, in the conditional case both full and reduced model parameters are required.

- **The spectral point-frequency estimator**

  The null hypothesis $H_0(\omega)$ for vanishing of $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\omega;\boldsymbol{\theta})$ (32) at the point frequency ω is $\Psi_{xy}(\omega) = 0$, where $\Psi(\omega)$ is the transfer function (24) for the VAR model, or (Geweke, 1982)

$$H_0(\omega) : \begin{cases} \displaystyle\sum_{k=1}^{p} A_{k,xy} \cos k\omega = 0 \\ \displaystyle\sum_{k=1}^{p} A_{k,xy} \sin k\omega = 0 \end{cases} \tag{115}$$

  For given ω, (115) represents $2n_x n_y$ constraints on the $pn_x n_y$ regression coefficient matrices[14] A_k. Calculation of the point-frequency asymptotic sampling distribution is in principle approachable via a similar technique as before (Proposition 3.1 must be adjusted for the case of more general linear constraints). However, we contend that in real-world applications it makes more sense in any case to consider inference on spectral Granger causality on a (possibly narrow-band) frequency range

[14] So if $\omega \neq k\pi$ for any k, then if $p \leq 2$ we have the original $H_0 : A_{k,xy} = 0$; *cf.* Section 3.3.

via the band-limited spectral GC $f_{\boldsymbol{Y}\to\boldsymbol{X}}(\mathcal{F};\boldsymbol{\theta})$ (33) as discussed in Section 3.2 rather than at point frequencies. Firstly, for a given VAR(p), if the "broadband" null condition H_0 is not satisfied, then the point-frequency null condition $H_0(\omega)$ will only be satisfied precisely at most at a finite number of (in practice unknown) point frequencies. Secondly, real-world spectral phenomena are likely to be to some extent broadband [e.g., power spectra of neural processes (Mitra and Bokil, 2008)] and/or otherwise blurred by noise.

The asymptotic sampling distribution of the point-frequency estimator is nonetheless at least of academic interest, since as far as the authors are aware, it remains unknown. Calculation of the point-frequency sampling distribution under the null hypothesis $H_0(\omega)$ is by no means intractable (at least in the unconditional case), although the computation of the Hessian is considerably complicated by the more restrictive null condition (115). The condition is, at least, linear, so may be readily transformed to conform to the preconditions for Proposition 3.1.

- **The single-regression F-test statistic**

  Granger causality may be tested using a standard F-test for vanishing of the appropriate regression coefficients (Hamilton, 1994). The population test statistic is

$$\frac{\operatorname{trace}\left[\Sigma^{\mathsf{R}}\right] - \operatorname{trace}[\Sigma_{xx}]}{\operatorname{trace}[\Sigma_{xx}]}\,, \tag{116}$$

  and the corresponding dual-regression estimator scaled by d_2/d_1 is asymptotically $F(d_1, d_2)$-distributed under the null hypothesis, where $d_1 = pn_xn_y$ and $d_2 = n_x[N - p(n_x + n_y + 1)]$ are the respective degrees of freedom. Anecdotally, (see, e.g., Hamilton, 1994; Lütkepohl, 2005), and especially for smaller sample lengths, this may yield a more powerful test than the (dual-regression) LR test. It would thus be of interest to investigate whether the single-regression F-statistic form (116) obtained by setting $\Sigma^{\mathsf{R}} = V(\boldsymbol{\theta})$ as in (38) similarly yields a more powerful test than the single-regression Geweke form (39). This statistic satisfies the conditions for Proposition 3.1, and the Hessian is as easily calculated as in Section 3.1.

  We remark that the statistic (116) is less useful as a quantitative measure of causal effect, as it lacks the information-theoretic interpretation of the Geweke form (31) (Barnett et al., 2009; Barnett and Bossomaier, 2013), nor is it invariant under as broad a set of transformations (Barrett et al., 2010), or under invertible digital filtering (Barnett and Seth, 2011).

- **The state-space Granger causality statistic**

  The state-space GC statistic, unconditional and conditional, in time and frequency domains, was introduced in Barnett and Seth (2015) (see also Solo, 2016), and is calculated from the innovations-form state-space parameters $\boldsymbol{\theta} = (A, C, K, \Sigma)$, for which reduced-model parameters A^{R} and C^{R} are known, while K^{R} and Σ^{R} appear as solutions of a certain DARE (*cf.* Section 3.1 and Appendix B). The state-space approach extends GC estimation and inference from the class of finite-order VAR models to the super-class of state-space (equivalently VARMA) models. The power of the method derives from the fact that (i) unlike for the class of finite-order VAR models, the class of state-space models is closed under subprocess extraction (an essential ingredient of the Granger causality construct), and (ii) many real-world data, notably econometric and neurophysiological, have a strong moving-average component, and are thus more parsimoniously represented as VARMA rather than pure VAR models. The class of state-space models is in addition—again in contrast to the finite order VAR class—closed under sub-sampling, temporal/spatial aggregation, observation noise, and digital filtering – all common features of real-world data acquisition and observation procedures.

  The single-regression GC statistic as defined here for a VAR(p) model is essentially the state-space GC statistic for a state-space model which describes the same stochastic process (Section 3.1 and

Appendix B). The state-space estimator may be defined in similar terms to the VAR case, i.e., based on ML estimates for the model parameters. In comparison with the VAR case, there are two challenges to derivation of the asymptotic null sampling distribution for the state-space GC estimator via the 2nd-order Delta Method (Proposition 3.1): (i) calculation of the Fisher information (i.e., distribution of the ML parameter estimators), and (ii) non-linearity of the null condition (Barnett and Seth, 2015, eq. 17). While (ii) complicates calculation of the Hessian, (i) is likely to present a more formidable obstacle, due to the considerable complexity of a closed-form expression for the Fisher information matrix (Klein et al., 2000).

5.3 The alternative hypothesis

We may consider two approaches to approximating the sampling distribution of the time-domain and band-limited spectral estimators, which address two distinct scenarios.

In the first scenario, we suppose given a fixed true parameter $\boldsymbol{\theta} \notin \Theta_0$, and consider the asymptotic sampling distribution of the GC statistic as sample length $N \to \infty$. In this case, the preconditions of Proposition 3.1 certainly do not apply; in particular, the gradient of the statistic will not in general vanish at $\boldsymbol{\theta}$, so that a 1st-order multivariate Delta Method is appropriate. This yields a normal distribution for the estimator, with mean equal to the population GC. If the statistic is $f\big(\hat{\boldsymbol{\theta}}\big)$, then explicitly we have

$$\sqrt{N}\left[f\big(\hat{\boldsymbol{\theta}}\big) - f(\boldsymbol{\theta})\right] \xrightarrow{d} \mathcal{N}\big(0, \nabla f(\boldsymbol{\theta}) \cdot \Omega(\boldsymbol{\theta}) \cdot \nabla f(\boldsymbol{\theta})^{\mathsf{T}}\big) \tag{117}$$

The variance

$$\sigma^2 = \nabla f(\boldsymbol{\theta}) \cdot \Omega(\boldsymbol{\theta}) \cdot \nabla f(\boldsymbol{\theta})^{\mathsf{T}} \tag{118}$$

may be computed from the known form of the statistic, although the gradients are harder to calculate, since (1) in the time domain the DARE does not, as in the null case (Section 3.1) collapse to a DLYAP (54), while (2) in the spectral band-limited case (Section 3.2), the transfer function $\Psi(\omega; \boldsymbol{\theta})$ is no longer block-triangular. Gradients, furthermore, must be calculated with respect to all (rather than just null) parameters.

This scenario is more pertinent in a realistic empirical situation where, for instance, we are reasonably confident (via a Projection Test as described in Section 4.1) that an estimated GC is significantly different from zero, and we would like to put confidence bounds on the estimate; it addresses the *efficiency* of the estimator.

Under the second—and more difficult to analyse—scenario, we suppose that sample length N is fixed (but large), and we consider the limiting distribution of the single-regression GC estimator as the true non-null parameter approaches the null subspace Θ_0. We are now in the territory of Wald (1943), where the asymptotics of the Taylor expansion (on which the 1st- and 2nd-order Delta Methods are based) become a "balancing act" between sample length N and the distance between the true parameter and the null subspace. This is likely to be difficult to calculate; we conjecture that (by analogy with Wald's Theorem) the asymptotic distribution will be a non-central generalised χ^2. This scenario is more pertinent to analysis of the *power* of the statistic.

5.4 Concluding remarks

In this study we analysed the hitherto unknown large-sample behaviour of single-regression Granger causality estimators. As quantitative measures of causal effect/information transfer, these estimators clearly outperform their (problematic) classical dual-regression likelihood-ratio counterparts, through consistency and reduced bias and variance. Studying their asymptotic null distributions is therefore of importance, as it admits the construction of novel and tailored hypothesis tests. We have shown the distributions to be generalised χ^2 in both the time-domain and band-limited spectral case, and obtained the distributional parameters in readily-calculated form. Based on these results, we introduced a novel asymptotically-valid

"Projection Test" of the null hypothesis of vanishing causality. This test should prove especially useful for testing Granger causality on specified frequency bands, a commonplace desideratum in various fields of application, including neuroimaging and econometric time-series analysis. We showed that the approach remains valid even if the model order is unknown, provided a consistent model order selection criterion is used. Our approach may be extended to the conditional case, and, potentially, beyond pure autoregressive modelling to causal inference based on the state-space Granger causality estimator. Finally, we outlined an approach to evaluation of the sampling distribution under the alternative hypothesis, knowledge of which would be useful for the analysis of statistical power and construction of confidence bounds. These extensions merit further detailed investigation, being highly relevant to a wide range of practical applications.

Acknowledgements

The authors would like to thank the Dr. Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science. We are also grateful to Anil K. Seth and Michael Wibral for useful comments on this work.

Author contributions

Authors Gutknecht and Barnett contributed equally to this work.

Appendices

A Proof of Proposition 3.1

Let $\boldsymbol{\theta} = [\boldsymbol{x}^\mathsf{T}\ \boldsymbol{y}^\mathsf{T}]^\mathsf{T}$ where $x_i = \theta_i$, $i = 1, \ldots, d$ and $y_j = \theta_{d+j}$, $j = 1, \ldots, s$. Since by definition $f(0, \boldsymbol{y}) = 0\ \forall \boldsymbol{y}$, we have immediately $\nabla_{\boldsymbol{y}} f(0, \boldsymbol{y}) = 0\ \forall \boldsymbol{y}$. Treating $\boldsymbol{y}$ as fixed, we expand $f(\boldsymbol{x}, \boldsymbol{y})$ in a Taylor series around $\boldsymbol{x} = 0$:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \nabla_{\boldsymbol{x}} f(0, \boldsymbol{y})\boldsymbol{x} + \tfrac{1}{2}\boldsymbol{x}^\mathsf{T}\nabla^2_{\boldsymbol{xx}} f(0, \boldsymbol{y})\boldsymbol{x} + \tfrac{1}{2}\boldsymbol{x}^\mathsf{T} K(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{x} \tag{119}$$

where for fixed $\boldsymbol{y}$, $K(\boldsymbol{x}, \boldsymbol{y})$ is a $d \times d$ matrix function of $\boldsymbol{x}$, and $\lim_{\boldsymbol{x}\to 0} \|K(\boldsymbol{x}, \boldsymbol{y})\| = 0$. Now we show that since $f(\boldsymbol{x}, \boldsymbol{y})$ is non-negative, we must have $\nabla_{\boldsymbol{x}} f(0, \boldsymbol{y}) = 0\ \forall \boldsymbol{y}$. Suppose, say, $\nabla_{x_1} f(0, \boldsymbol{y}) = -g < 0$. Setting $x_1 = \varepsilon > 0$ [if $\nabla_{x_1} f(0, \boldsymbol{y}) > 0$ we take $x_1 = -\varepsilon$] and $x_2 = \ldots = x_d = 0$, (119) yields

$$f(\boldsymbol{x}, \boldsymbol{y}) = -g\varepsilon + \tfrac{1}{2}\left[\nabla^2_{x_1 x_1} f(0, \boldsymbol{y}) + K_{11}(\boldsymbol{x}, \boldsymbol{y})\right]\varepsilon^2 \tag{120}$$

Now since $\lim_{\varepsilon\to 0} \|K(\boldsymbol{x}, \boldsymbol{y})\| = 0$, we can always choose ε small enough that

$$\tfrac{1}{2}\left[\nabla^2_{x_1 x_1} f(0, \boldsymbol{y}) + K_{11}(\boldsymbol{x}, \boldsymbol{y})\right]\varepsilon < g \tag{121}$$

so that $f(\varepsilon, 0, ..., 0, \boldsymbol{y}) < 0$, a contradiction. Thus we have $\nabla_{\boldsymbol{x}} f(0, \boldsymbol{y}) = 0\ \forall \boldsymbol{y}$, proving 3.1a.

From (119) we thus have

$$f(\boldsymbol{x}, \boldsymbol{y}) = \tfrac{1}{2}\boldsymbol{x}^\mathsf{T}\nabla^2_{\boldsymbol{xx}} f(0, \boldsymbol{y})\boldsymbol{x} + \tfrac{1}{2}\boldsymbol{x}^\mathsf{T} K(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{x} \tag{122}$$

To see that $\nabla^2_{\boldsymbol{xx}} f(0, \boldsymbol{y})$ must be positive-semidefinite, we assume the contrary. We may then find a unit d-dimensional vector $\boldsymbol{u}$ such that $\boldsymbol{u}^\mathsf{T}\nabla^2_{\boldsymbol{xx}} f(0, \boldsymbol{y})\boldsymbol{u} = -G < 0$. Setting $\boldsymbol{x} = \varepsilon\boldsymbol{u}$, we may then choose ε small enough that $\boldsymbol{u}^\mathsf{T} K(\varepsilon\boldsymbol{u}, \boldsymbol{y})\boldsymbol{u} < G$, so that again $f(\boldsymbol{x}, \boldsymbol{y})$ is negative and we have a contradiction. Finally, $\nabla^2_{\boldsymbol{xy}} f(0, \boldsymbol{y}) = \nabla^2_{\boldsymbol{yy}} f(0, \boldsymbol{y}) = 0\ \forall \boldsymbol{y}$ follows directly from (122), and we have established 3.1b.

We now prove 3.1c using a 2nd-order Delta Method (Lehmann and Romano, 2005). Let $\boldsymbol{\theta} \in \Theta_0$. Since $f(\boldsymbol{\theta})$ and its gradient $\nabla f(\boldsymbol{\theta})$ both vanish, the Taylor expansion of $f(\boldsymbol{\vartheta}_N)$ around $\boldsymbol{\theta}$ takes the form

$$f(\boldsymbol{\vartheta}_N) = \tfrac{1}{2} (\boldsymbol{\vartheta}_N - \boldsymbol{\theta})^\mathsf{T} \mathbf{H}(\boldsymbol{\theta}) (\boldsymbol{\vartheta}_N - \boldsymbol{\theta}) + (\boldsymbol{\vartheta}_N - \boldsymbol{\theta})^\mathsf{T} K(\boldsymbol{\vartheta}_N) (\boldsymbol{\vartheta}_N - \boldsymbol{\theta}) \tag{123}$$

where $\mathbf{H}(\boldsymbol{\theta}) = \nabla^2 f(\boldsymbol{\theta})$ is the Hessian matrix of f evaluated at $\boldsymbol{\theta}$, and $\lim_{\boldsymbol{\theta}' \to \boldsymbol{\theta}} K(\boldsymbol{\theta}') = 0$. Multiplying both sides by the sample size N, we have

$$N f(\boldsymbol{\vartheta}_N) = \tfrac{1}{2} \left[\sqrt{N} (\boldsymbol{\vartheta}_N - \boldsymbol{\theta})\right]^\mathsf{T} \mathbf{H}(\boldsymbol{\theta}) \left[\sqrt{N} (\boldsymbol{\vartheta}_N - \boldsymbol{\theta})\right] + \left[\sqrt{N} (\boldsymbol{\vartheta}_N - \boldsymbol{\theta})\right]^\mathsf{T} K(\boldsymbol{\vartheta}_N) \left[\sqrt{N} (\boldsymbol{\vartheta}_N - \boldsymbol{\theta})\right] \tag{124}$$

But by assumption $\sqrt{N}(\boldsymbol{\vartheta}_N - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{Z}$ as $N \to \infty$, where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \Omega)$. Thus, by the CMT (Van der Vaart, 2000), we have

$$N f(\boldsymbol{\vartheta}_N) \xrightarrow{d} \tfrac{1}{2} \boldsymbol{Z}^\mathsf{T} \mathbf{H}(\boldsymbol{\theta}) \boldsymbol{Z} \tag{125}$$

as $N \to \infty$, and 3.1c follows immediately from (125) and 3.1b.

B State-space solution for the reduced VAR model parameters

Following Barnett and Seth (2015), given a VAR(p) model (35) with parameters $\boldsymbol{\theta} = (A; \Sigma)$, with $A = [A_1\ A_2\ \ldots\ A_p]$, we create an equivalent "innovations form" state-space model (Hannan and Deistler, 2012)

$$\boldsymbol{Z}_{t+1} = \mathbf{A} \boldsymbol{Z}_t + K \boldsymbol{\varepsilon}_t \tag{126a}$$
$$\boldsymbol{U}_t = C \boldsymbol{Z}_t + \quad \boldsymbol{\varepsilon}_t \tag{126b}$$

where

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \ldots & A_{p-1} & A_p \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix} \qquad C = \begin{bmatrix} A_1 & A_2 & \ldots & A_{p-1} & A_p \end{bmatrix} \qquad K = \begin{bmatrix} I \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{127}$$

$\mathbf{A}$ is the $pn \times pn$ state transition matrix [the companion matrix (18) for the VAR(p) (35)], K the $pn \times n$ Kalman gain matrix, and C the $n \times pn$ observation matrix. As before, we use notation x for the indices $1, \ldots, n_x$, y for the indices $n_x + 1, \ldots, n$ and (after the the Matlab® convention) a colon to donate "all indices". The subprocess $\boldsymbol{X}_t$ then satisfies the state-space model

$$\boldsymbol{Z}_{t+1} = \mathbf{A} \boldsymbol{Z}_t \ + K \boldsymbol{\varepsilon}_t \tag{128a}$$
$$\boldsymbol{X}_t = C^\mathsf{R} \boldsymbol{Z}_t + \quad \boldsymbol{\varepsilon}_{x,t} \tag{128b}$$

with $C^\mathsf{R} = C_{x:}$. This state-space model is no longer in innovations form; we can, however (see Barnett and Seth, 2015) derive an innovations-form state-space model for $\boldsymbol{X}_t$ by solving the discrete-time algebraic Riccati equation (DARE)

$$\begin{aligned} P - \mathbf{A} P \mathbf{A}^\mathsf{T} &= Q - K^\mathsf{R} \Sigma^\mathsf{R} K^{\mathsf{RT}} \\ \Sigma^\mathsf{R} &= C^\mathsf{R} P C^{\mathsf{RT}} + R \\ K^\mathsf{R} &= \left(\mathbf{A} P C^{\mathsf{RT}} + S\right) \left[\Sigma^\mathsf{R}\right]^{-1} \end{aligned} \tag{129}$$

with

$$Q = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \qquad S = \begin{bmatrix} \Sigma_{:x} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad R = \Sigma_{xx} \tag{130}$$

which are, respectively, $pn \times pn$, $pn \times n_x$ and $n_x \times n_x$. Then

$$\boldsymbol{Z}_{t+1} = \mathbf{A}\boldsymbol{Z}_t + K^{\mathsf{R}}\boldsymbol{\varepsilon}_t^{\mathsf{R}} \tag{131a}$$

$$\boldsymbol{X}_t = C^{\mathsf{R}}\boldsymbol{Z}_t + \boldsymbol{\varepsilon}_t^{\mathsf{R}} \tag{131b}$$

is in innovations form [note that the $\boldsymbol{\varepsilon}_t^{\mathsf{R}}$ are precisely the reduced residuals in (29)] and we may solve the DARE for Σ^{R}, which implicitly defines the mapping (38) $\Sigma^{\mathsf{R}} = V(\boldsymbol{\theta})$ which is required for the single-regression GC statistic $F^{\mathsf{SR}}_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$ (39).

We may, in fact, confirm that $K^{\mathsf{R}}, \Sigma^{\mathsf{R}}$ are solutions of the lower-dimensional DARE

$$\begin{aligned} \boldsymbol{P} - \boldsymbol{A}_{yy}\boldsymbol{P}\boldsymbol{A}_{yy}^{\mathsf{T}} &= \boldsymbol{\Sigma}_{yy} - K^{\mathsf{R}}\Sigma^{\mathsf{R}}K^{\mathsf{RT}} \\ \Sigma^{\mathsf{R}} &= A_{xy}\boldsymbol{P}A_{xy}^{\mathsf{T}} + \Sigma_{xx} \\ K^{\mathsf{R}} &= \left(\boldsymbol{A}_{yy}\boldsymbol{P}A_{xy}^{\mathsf{T}} + \boldsymbol{\Sigma}_{yx}\right)\left[\Sigma^{\mathsf{R}}\right]^{-1} \end{aligned} \tag{132}$$

where $\boldsymbol{P}$ is $pn_y \times pn_y$,

$$\boldsymbol{A}_{yy} = \begin{bmatrix} A_{1,yy} & A_{2,yy} & \dots & A_{p-1,yy} & A_{p,yy} \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix} \qquad A_{xy} = \begin{bmatrix} A_{1,xy} & A_{2,xy} & \dots & A_{p-1,xy} & A_{p,xy} \end{bmatrix} \tag{133}$$

which are, respectively, $pn_y \times pn_y$ and $n_x \times pn_y$, and

$$\boldsymbol{\Sigma}_{yy} = \begin{bmatrix} \Sigma_{yy} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{yx} = \begin{bmatrix} \Sigma_{yx} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{134}$$

respectively, $pn_y \times pn_y$, $pn_y \times n_x$. To see this, we note that under our assumptions, the solution of the DARE (129) exists, and is unique (Solo, 2016). Then, setting

$$P_{kl} = \begin{bmatrix} 0_{n_x \times n_x} & 0_{n_x \times n_y} \\ 0_{n_y \times n_x} & \boldsymbol{P}_{kl} \end{bmatrix} \tag{135}$$

we may verify that

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1p} \\ \vdots & & \vdots \\ P_{p1} & \cdots & P_{pp} \end{bmatrix} \tag{136}$$

solves the original DARE (129) if

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{11} & \cdots & \boldsymbol{P}_{1p} \\ \vdots & & \vdots \\ \boldsymbol{P}_{p1} & \cdots & \boldsymbol{P}_{pp} \end{bmatrix} \tag{137}$$

solves the reduced-dimension DARE (132).

C Granger causality for a general bivariate VAR(1) process

Consider the bivariate VAR(1)

$$X_t = a_{xx}X_{t-1} + a_{xy}Y_{t-1} + \varepsilon_{xt} \tag{138a}$$
$$Y_t = a_{yx}X_{t-1} + a_{yy}Y_{t-1} + \varepsilon_{yt} \tag{138b}$$

with parameters

$$A = \begin{bmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{bmatrix}, \qquad \Sigma = \mathbf{E}\big[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^{\mathsf{T}}\big] = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \tag{139}$$

The transfer function is then $\Psi(\omega) = \Phi(\omega)^{-1}$ with $\Phi(\omega) = I - Az$, and we have the spectral factorisation (28) of the CPSD $S(\omega)$ of the VAR process, which holds for z on the unit circle $|z| = 1$ in the complex plane. Setting $\Delta(\omega) = |\Phi(\omega)|$ (determinant), we have

$$\Psi(\omega) = \begin{bmatrix} 1 - a_{xx}z & -a_{xy}z \\ -a_{yx}z & 1 - a_{yy}z \end{bmatrix}^{-1} = \Delta(\omega)^{-1} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ a_{yx}z & 1 - a_{xx}z \end{bmatrix} \tag{140}$$

This leads to

$$S(\omega) = |\Delta(\omega)|^{-2} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ a_{yx}z & 1 - a_{xx}z \end{bmatrix} \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \begin{bmatrix} 1 - a_{yy}\bar{z} & a_{yx}\bar{z} \\ a_{xy}\bar{z} & 1 - a_{xx}\bar{z} \end{bmatrix} \tag{141}$$

on $z = 1$, where $\bar{z}$ is the complex conjugate of z; here, for a complex variable w, $|w|$ denotes the norm $\sqrt{w\bar{w}}$.

We wish to calculate the GC $F_{Y\to X}$. If v is the residuals variance for the VAR representation of the subprocess X_t, then the GC is just $F_{Y\to X} = \log v - \log \sigma_{xx}$. To solve for v we could use the state-space method of Barnett and Seth (2015), but here we use an explicit spectral factorisation for the CPSD $S_{xx}(\omega)$ of X_t.

Let $\psi(\omega)$ be the transfer function of the process X_t. It then satisfies the spectral factorisation

$$S_{xx}(\omega) = v|\psi(\omega)|^2 \tag{142}$$

with $\psi(0) = 1$. We may now calculate (we denote terms we don't need by "$\cdots$").

$$\begin{aligned} S(\omega) &= |\Delta(\omega)|^{-2} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \begin{bmatrix} 1 - a_{yy}\bar{z} & \cdots \\ a_{xy}\bar{z} & \cdots \end{bmatrix} \\ &= |\Delta(\omega)|^{-2} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} \sigma_{xx}(1 - a_{yy}\bar{z}) + \sigma_{xy}a_{xy}\bar{z} & \cdots \\ \sigma_{yx}(1 - a_{yy}\bar{z}) + \sigma_{yy}a_{xy}\bar{z} & \cdots \end{bmatrix} \end{aligned}$$

We now calculate [with $z = e^{-i\omega}$ so that $\mathfrak{Re}\{z\} = \frac{1}{2}(z + \bar{z}) = \cos\omega$]:

$$\begin{aligned} S_{xx}(\omega) &= |\Delta(\omega)|^{-2} \left\{(1 - a_{yy}z)[\sigma_{xx}(1 - a_{yy}\bar{z}) + \sigma_{xy}a_{xy}\bar{z})] + a_{xy}z[\sigma_{yx}(1 - a_{yy}\bar{z}) + \sigma_{yy}a_{xy}\bar{z})]\right\} \\ &= |\Delta(\omega)|^{-2} \left\{\sigma_{xx}|1 - a_{yy}z|^2 + \sigma_{xy}a_{xy}[(1 - a_{yy}z)\bar{z} + (1 - a_{yy}\bar{z})z] + \sigma_{yy}a_{xy}^2 z\bar{z}\right\} \\ &= |\Delta(\omega)|^{-2} \left\{\sigma_{xx}[1 - a_{yy}(z + \bar{z}) + a_{yy}^2] + \sigma_{xy}a_{xy}(z + \bar{z} - 2a_{yy}) + \sigma_{yy}a_{xy}^2\right\} \\ &= |\Delta(\omega)|^{-2} \left\{\sigma_{xx}[1 - 2a_{yy}\cos\omega + a_{yy}^2] + 2\sigma_{xy}a_{xy}(\cos\omega - a_{yy}) + \sigma_{yy}a_{xy}^2\right\} \end{aligned}$$

and finally

$$S_{xx}(\omega) = |\Delta(\omega)|^{-2}(P - Q\cos\omega) \tag{143}$$

where we have set

$$P = \sigma_{xx}(1 + a_{yy}^2) - 2\sigma_{xy}a_{xy}a_{yy} + \sigma_{yy}a_{xy}^2 \tag{144a}$$
$$Q = 2(\sigma_{xx}a_{yy} - \sigma_{xy}a_{xy}) \tag{144b}$$

The form of this expression suggests that the transfer function $\psi(\omega)$ should take the form

$$\psi(\omega) = \Delta(\omega)^{-1}(1 - bz) \tag{145}$$

for some constant b. Note that this implies that X_t is ARMA(2,1). Then $|\psi(\omega)|^2 = |\Delta(\omega)|^{-2}(1 + b^2 - 2b\cos\omega)$ and the spectral factorisation $S_{xx}(\omega) = v|\psi(\omega)|^2$ now reads:

$$v(1 + b^2 - 2b\cos\omega) = P - Q\cos\omega \tag{146}$$

This must hold for at point on the unit circle—i.e., for all ω—so we must have

$$v(1 + b^2) = P \tag{147}$$
$$vb = \tfrac{1}{2}Q \tag{148}$$

We may now solve for v. The second equation gives $v^2b^2 = \frac{1}{4}Q^2$, so multiplying the first equation through by v we obtain the quadratic equation for v:

$$v^2 - Pv + \tfrac{1}{4}Q^2 = 0 \tag{149}$$

with solutions

$$v = \frac{1}{2}\left(P \pm \sqrt{P^2 - Q^2}\right) \tag{150}$$

We need to take the "+" solution, as this yields the correct (zero) result for the null case $a_{xy} = 0$. Note that only the $Y \to X$ "causal" coefficient a_{xy} and the Y autoregressive coefficient a_{yy} appear in the expression for $F_{Y\to X}$.

From (32), the spectral GC from $Y \to X$ is

$$f_{Y\to X}(\omega) = \log\frac{P - Q\cos\omega}{P - Q\cos\omega - a_{xy}^2\sigma_{yy|x}} \tag{151}$$

where

$$\sigma_{yy|x} = \sigma_{yy} - \frac{\sigma_{xy}^2}{\sigma_{xx}} = \sigma_{yy}\left(1 - \kappa^2\right) \tag{152}$$

with $\kappa = \dfrac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}$ the residuals correlation.

For the sampling distributions, we shall also need the (inverse of) the covariance matrix Γ_0 of the process $[X_t^\mathsf{T}\ Y_t^\mathsf{T}]^\mathsf{T}$ on the null space $a_{xy} = 0$. Solving the DLYAP equation $\Gamma_0 - A\Gamma_0A^\mathsf{T} = \Sigma$ for $\Gamma_0 = \begin{bmatrix} p & r \\ r & q \end{bmatrix}$ yields

$$p = \left(1 - a_{xx}^2\right)^{-1}\sigma_{xx} \tag{153a}$$
$$r = (1 - a_{xx}a_{yy})^{-1}\left[\sigma_{xy} + a_{xx}a_{yx}\left(1 - a_{xx}^2\right)^{-1}\sigma_{xx}\right] \tag{153b}$$
$$q = \left(1 - a_{yy}^2\right)^{-1}\left[\sigma_{yy} + 2a_{yy}a_{yx}(1 - a_{xx}a_{yy})^{-1}\sigma_{xy} + a_{yx}^2(1 + a_{xx}a_{yy})\left(1 - a_{xx}^2\right)^{-1}(1 - a_{xx}a_{yy})^{-1}\sigma_{xx}\right] \tag{153c}$$

and we have in particular

$$\omega_{yy} = [\Gamma_0^{-1}]_{yy} = \frac{p}{pq - r^2} \tag{154}$$

Note also that in the null case $a_{xy} = 0$, the spectral radius is $\rho = \max(|a_{xx}|, |a_{yy}|)$.

D Parametrised sampling of the VAR model space

Consider, for given number of variables n and model order p, the parameter space $\Theta = \{(A, \Sigma) : A \text{ is } n \times pn \text{ with } \rho(A) < 1, \Sigma \text{ is } n \times n \text{ positive-definite}\}$ of VAR(p) models. Firstly, we note that the residuals covariance matrix Σ can be taken to be a *correlation* matrix; this can always be achieved by a rescaling of variables leaving Granger causalities invariant. Further GC invariances under linear transformation of variables (Barrett et al., 2010) allow further effective dimensional reduction of Θ; however, even under these general transformations, and under the constraint $\rho(A) < 1$, the quotient space of Θ has infinite Lebesgue measure[15]; thus we cannot generate uniform variates (it is questionable whether this would in any case be appropriate to a given empirical scenario). Here we utilise a practical and flexible scheme for generation of variates on Θ, parametrised by spectral radius ρ, log-generalised correlation[16] $\gamma = -\log|\Sigma| + \sum_i \log \Sigma_{ii}$, and population GC $F = F_{\boldsymbol{Y}\to\boldsymbol{X}}(\boldsymbol{\theta})$, all of which have a critical impact on GC sampling distributions.

To generate a random correlation matrix Σ of dimension n with given generalised correlation γ, we use the following algorithm:

1. Starting with an $n \times n$ matrix with components iid $\sim \mathcal{N}(0,1)$, we compute its QR-decomposition $[Q, R]$. The matrix $M_{ij} = Q_{ij} \cdot \text{sign}(R_{jj})$ is then a random orthogonal matrix.

2. Next we create a random n-dimensional variance vector $\boldsymbol{v}$ with components v_i iid $\sim \chi^2(1)$. The matrix $V = M \cdot \text{diag}(\boldsymbol{v}) \cdot M^\mathsf{T}$ is then positive-definite, and for the corresponding correlation matrix $\Sigma_{ij} = V_{ij}/\sqrt{V_{ii}V_{jj}}$ we have $\gamma^* = -\sum_i \log v_i + \sum_i \log V_{ii}$.

   If necessary, we repeat steps 1,2 until $\gamma^* \geq \gamma$ (this may fail if γ is too large).

3. Using a binary chop, we find a constant c such that, iteratively replacing $\boldsymbol{v} \leftarrow \boldsymbol{v} + c$, γ^* falls within an acceptable tolerance of γ (this generally converges). The correlation matrix Σ is then returned,

For a VAR coefficients matrix sequence $A = [A_1\ A_2\ \ldots\ A_p]$, the spectral radius $\rho(A)$ is given by (20). If λ is a constant, it is easy to show that if $\tilde{A}$ is the sequence $[\lambda A_1\ \lambda^2 A_2\ \ldots\ \lambda^p A_p]$, then $\rho(\tilde{A}) = \lambda\rho(A)$. Thus any VAR coefficients sequence may be exponentially weighted so that its spectral radius takes a given value. Such weighting, however, has the side-effect of exponential decay of the A_k with lag k, which is, anecdotally, unrealistic[17]. We observe empirically that we can compensate for this decay reasonably consistently across number of variables and model orders by scaling all coefficients by A_k by $e^{-\sqrt{p}w}$ for some constant w; here we choose $w = 1$, which generally achieves a more realistic gradual and approximately linear decay. To generate a random VAR model with given generalised correlation γ and given spectral radius ρ, our procedure is as follows:

1. We generate a random correlation matrix Σ with generalised correlation γ as described above.

2. We generate p $n \times n$ coefficient matrices A_k with components iid $\sim \mathcal{N}(0,1)$. The A_k are the weighted uniformly by $e^{-\sqrt{p}w}$.

To enforce the null condition $A_{k,xy} = 0$,

3. The $A_{k,xy}$ components are all set to zero.

4. The A_k coefficients sequence is scaled exponentially by an appropriate constant λ, so as to achieve the given spectral radius ρ.

To instead enforce a given (non-null) population GC value F,

[15] Although the space of $n \times n$ correlation matrices has finite measure.

[16] For Gaussian covariance matrices, log-generalised correlation coincides with *multi-information* (Studený and Vejnarová, 1998). If $R = (\rho_{ij})$ is a correlation matrix with all $\rho_{ij} \ll 1$ for $i \neq j$, then $-\log|R| \approx \sum_{i<j} \rho_{ij}^2$.

[17] At least, in the authors' experience, for neural or econometric data.

3. The $A_{k,xy}$ components are scaled uniformly by a constant c.

4. The A_k coefficients sequence is scaled exponentially by an appropriate constant λ, so as to achieve the given spectral radius ρ.

   Under steps $3, 4$ the population GC depends monotonically on c; consequently,

5. We perform a binary chop on c, iterating steps $3, 4$ until the GC is within an acceptable tolerance of F (this generally converges quickly).

In all simulations except for the bivariate model (Section 3.3), we used $\gamma = 1$; spectral radii and population GC values are as indicated in the plots. Convergence tolerances were set to[18] $\sqrt{\text{machine } \varepsilon}$.

References

Amari, S. (2016). *Information Geometry and its Applications*. Springer, Japan.

Barnett, L., Barrett, A. B., and Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.*, 103(23):0238701.

Barnett, L., Barrett, A. B., and Seth, A. K. (2018a). Misunderstandings regarding the application of Granger causality in neuroscience. *Proc. Natl. Acad. Sci. USA*, 115(29):E6676–E6677.

Barnett, L., Barrett, A. B., and Seth, A. K. (2018b). Solved problems for Granger causality in neuroscience: A response to Stokes and Purdon. *NeuroImage*, 178:744–748.

Barnett, L. and Bossomaier, T. (2013). Transfer entropy as a log-likelihood ratio. *Phys. Rev. Lett.*, 109(13):0138105.

Barnett, L. and Seth, A. K. (2011). Behaviour of Granger causality under filtering: Theoretical invariance and practical application. *J. Neurosci. Methods*, 201(2):404–419.

Barnett, L. and Seth, A. K. (2014). The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *J. Neurosci. Methods*, 223:50–68.

Barnett, L. and Seth, A. K. (2015). Granger causality for state-space models. *Phys. Rev. E (Rapid Communications)*, 91(4):040101(R).

Barrett, A. B., Barnett, L., and Seth, A. K. (2010). Multivariate Granger causality and generalized variance. *Phys. Rev. E*, 81(4):041907.

Chen, Y., Bressler, S. L., and Ding, M. (2006). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *J. Neuro. Methods*, 150:228–237.

Dhamala, M., Liang, H., Bressler, S. L., and Ding, M. (2018). Granger-Geweke causality: Estimation and interpretation. *NeuroImage*, 175:460–463.

Dhamala, M., Rangarajan, G., and Ding, M. (2008a). Analyzing information flow in brain networks with nonparametric Granger causality. *Neuroimage*, 41:354–362.

Dhamala, M., Rangarajan, G., and Ding, M. (2008b). Estimating Granger causality from Fourier and wavelet transforms of time series data. *Phys. Rev. Lett.*, 100:018701.

[18] Approximately 1.5×10^{-8} under the IEEE 754-2008 binary64 floating point standard.

Ding, M., Chen, Y., and Bressler, S. L. (2006). Granger causality: Basic theory and application to neuroscience. In Schelter, B., Winterhalder, M., and Timmer, J., editors, *Handbook of Time Series Analysis*, pages 437–460. Wiley-VCH Verlag GmbH & Co. KGaA.

Doob, J. (1953). *Stochastic Processes*. John Wiley, New York.

Faes, L., Stramaglia, S., and Marinazzo, D. (2017). On the interpretability and computational reliability of frequency-domain Granger causality. *F1000Research*, 6(1710). Version 1; Referees: 2 approved.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.*, 77(378):304–313.

Geweke, J. (1984). Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.*, 79(388):907–915.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Hannan, E. J. and Deistler, M. (2012). *The Statistical Theory of Linear Systems*. SIAM, Philadelphia, PA, USA.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Stat. Soc. B Met.*, 41(2):190–195.

Jones, D. A. (1983). Statistical analysis of empirical models fitted by optimization. *Biometrika*, 70(1):67–88.

Klein, A., Mélard, G., and Zahaf, T. (2000). Construction of the exact Fisher information matrix of Gaussian time series models by means of matrix differential rules. *Linear Algebra Appl.*, 321(1):209–232. Eighth Special Issue on Linear Algebra and Statistics.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Science+Business Media, LLC, New York, NY, USA, 3rd edition.

Lütkepohl, H. (1993). Testing for causation between two variables in higher dimensional VAR models. In Schneeweiß, H. and Zimmerman, K., editors, *Studies in Applied Econometrics*, pages 75–91. Physica-Verlag HD, Heidelberg.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.

Masani, P. (1966). Recent trends in multivariate prediction theory. In Krishnaiah, P. R., editor, *Multivariate Analysis*, pages 351–382. Academic Press, New York.

McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific Publishing, Singapore.

Mitra, P. P. and Bokil, H. (2008). *Observed Brain Dynamics*. Oxford University Press, New York.

Mohsenipour, A. A. (2012). *On the Distribution of Quadratic Expressions in Various Types of Random Vectors*. PhD thesis, The University of Western Ontario, Electronic Thesis and Dissertation Repository, 955. https://ir.lib.uwo.ca/etd/955.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. A*, 231:289–337.

Rozanov, Y. A. (1967). *Stationary Random Processes*. Holden-Day, San Francisco.

Solo, V. (2016). State-space analysis of Granger-Geweke causality measures with application to fMRI. *Neural Comput.*, 28(5):914–949.

Stokes, P. A. and Purdon, P. L. (2017). A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proc. Natl. Acad. Sci. USA*, 114(34):7063–7072.

Stokes, P. A. and Purdon, P. L. (2018). Correction for Stokes and Purdon, A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proc. Natl. Acad. Sci. USA*, 115(29):E6964–E6964.

Studený, M. and Vejnarová, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 261–297. Springer Netherlands, Dordrecht.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *T. Am. Math. Soc.*, 54(3):426–482.

Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50(1,2):129–134.

Wiener, N. and Masani, P. (1957). The prediction theory of multivariate stochastic processes: I. The regularity condition. *Acta Math.*, 98:111–150.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24:471–494.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 6(1):60–62.

Wilson, G. T. (1972). The factorization of matricial spectral densities. *SIAM J. Appl. Math.*, 23(4):420–426.

Text width × height = 506.17488pt × 675.46852pt