Predictive Testing for Granger Causality via Posterior Simulation and Cross Validation

Gary J. Cornwall[1] Jeffrey A. Mills[2]
Beau A. Sauley[2]
Huibin Weng[2]

[1]Bureau of Economic Analysis
[2]Department of Economics, University of Cincinnati

November 18, 2018

Abstract

This paper develops a predictive approach to Granger causality testing that utilizes k-fold cross-validation and posterior simulation to perform out-of-sample testing. A Monte Carlo study indicates that the cross-validation predictive procedure has improved power in comparison to previously available out-of-sample testing procedures, matching the performance of the in-sample F-test while retaining the credibility of post sample inference. An empirical application to the Phillips curve is provided evaluating the evidence on Granger causality between inflation and unemployment rates.

Key words: Granger causality, predictive testing, posterior simulation, cross validation, out of sample testing.

Corresponding author: Jeff Mills, jeffrey.mills@uc.edu, 513.556.2619

1 Introduction

Granger causality testing is a standard procedure for analyzing multivariate time series that has seen widespread use across several disciplines, including economics (Yu et al., 2015; Ghysels et al., 2016; Diks and Wolski, 2016), physics (Dhamala et al., 2008; Barnett et al., 2009; Zhang et al., 2011; Attanasio et al., 2012; Barnett and Seth, 2015), and, more recently, neuroscience (Roebroeck et al., 2005, 2011; Liao et al., 2011; Hu et al., 2016; Barnett et al., 2017; Stokes and Purdon, 2017). Since Ashley et al. (1980), there has been considerable interest in developing tests that provide out-of-sample evaluation of evidence for Granger causality (Diebold and Mariano, 1995; Clark and McCracken, 2001, 2005, 2006; McCracken, 2007; Ashley and Tsang, 2014).

A common thread of the out-of-sample testing literature is that credible Granger causality testing "must rely primarily on the out-of-sample forecasting performance of models relating the original (non-prewhitened) series of interest." (Ashley et al., 1980). This follows Feigl's definition of "causality as predictability according to a law" so that Granger causality can be viewed as providing an evidential step useful for identifying causal relationships (Poirier, 1988). Further, Clark (2004) provides simulation evidence that out-of-sample test procedures avoid the possible spurious results due to overfitting that can arise with in-sample tests. On the other hand, as Inoue and Kilian (2005) point out, there is a loss of efficiency and power with out-of-sample tests due to partitioning the data and only using a sub-sample for estimation. This loss of power explains the continued popularity of the in-sample F-test (Sims, 1972), despite the potential for overfitting and pre-test estimator bias.

This paper develops an out-of-sample Granger causality test with empirical power characteristics close to that of the in-sample F-test, and superior to other out-of-sample tests. It has been shown that, under some conditions, the out-of-sample tests can produce greater power than the in-sample F-test, e.g. when discrete structural breaks are present in the time series (Chen, 2005b). For the majority of data generating processes considered in simulation studies however, the in-sample test provides a substantial increase in power compared to any of the out-of-sample tests currently available, particularly with small samples. Simulation results indicate that even under conditions most favorable to the in-sample F-test (stationary, stable data generating process with i.i.d. errors) the new testing procedure proposed herein has power close to that of the in-sample F-test. This is achieved by extending the ideas of Ashley and Tsang (2014), who employ a cross-validation approach to enable out-of-sample evaluation with modest sample sizes, resulting in a test procedure that consistently exhibits superior power to previously developed out-of-sample tests.

The main contributions of this paper are twofold. First, an alternative approach to cross-validation that is popular in the statistics literature, namely k-fold cross-validation and when $k = 1$ leave-one-out cross validation (LOO-CV), is considered (Stone, 1977; Picard and Cook, 1984; Arlot et al., 2010). This provides a substantial improvement in the power of the test, and eliminates the need to make an arbitrary choice of partition point. The LOO-CV approach has the benefits of an out-of-sample test (avoidance of spurious results due to overfitting and pre-test estimator bias), while the loss in efficiency and power is minimal since only one observation is omitted for estimation purposes.

Second, Bayesian predictive inference and MCMC sampling methods are employed to obtain the posterior predictive distribution of the proposed test statistic. The out-of-sample evaluation using cross validation employed by Ashley and Tsang (2014) results in a statistic, such as a root mean square prediction error (RMSPE) or an F-statistic, with unknown distribution. The out-of-sample "F-statistic" is no longer F-distributed and can be negative because the restricted sum of squares for out-of-sample predictions can be smaller than the unrestricted sum of squares if the restrictions lead to greater predictive accuracy. The approach thus requires asymptotic assumptions or bootstrapping. This paper explores MCMC posterior simulation methods for inference as an alternative. This allows evaluation of the exact posterior density for the test statistic (such as RMSPE) with any sample size and, by adopting the testing procedure developed in Mills (2018), allows computation of posterior odds ratios to compare out-of-sample predictive performance.

The new testing procedure is relatively easy to use in practice and not overly computationally burdensome. To provide some guidance to applied researchers, an example application with some further details on implementation is provided in section 4.[1]

The remainder of the paper is organized as follows. The proposed procedure is developed in section 2. Results of a Monte Carlo simulations examining the efficacy of the proposed procedure are given in section 3. Section

[1] Matlab, R and Julia code to implement the testing procedure is available at: https://github.com/tszanalytics/GrangerCausality

4 provides an empirical example related to the Phillips curve. Section 5 concludes.

2 Testing for Granger Causality via predictive cross validation

The series $x_{1:T}$ is said to Granger cause $y_{1:T}$ if past values of x_t, $x_{1:t-1}$, have additional power in forecasting y_t after controlling for the past of y_t, $y_{1:t-1}$ (Seth, 2007), so that $p(y_t|y_{1:t-1}, x_{1:t-1}) \neq p(y_t|y_{1:t-1})$. Testing for Granger causality (GC) generally takes place in a linear vector autoregression (VAR) model with Gaussian errors. The linear and Gaussian stochastic process assumptions can be relaxed in the following without great difficulty, but will be adhered to herein. Also, we focus on a bivariate model for simplicity of exposition, but the extension to more than two variables is straightforward. The testing framework is then a VAR(p),

$$\begin{bmatrix} \alpha(B) & \phi(B) \\ \beta(B) & \gamma(B) \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim \text{MVN}(\mathbf{0}, V), \tag{1}$$

where $\alpha(B), \phi(B), \beta(B), \gamma(B)$ are pth order polynomials in B, the backshift operator, $Bx_t = x_{t-1}$, i.e., $\alpha(B) = 1 - \alpha_0 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p$, $\phi(B) = -\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, etc., and MVN is a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix V, which is assumed to be iid homoskedastic. Homoskedasticity and serial independence of the error terms are imposed for expositional convenience; a more general heteroskedastic and time dependent error covariance structures can be modeled using MCMC in a seemingly unrelated regression framework (Mills and Namavari, 2017). Simulation results with a heteroskedastic data generating process (DGP) are provided in section 3.3.

For equation (1), the null hypothesis that x_t does not Granger cause y_t can be expressed as $H_0 : \phi_1 = ... = \phi_p = 0$, which naturally suggests the in-sample F-test widely used in applied research (Sims, 1972). As mentioned above, the drawback with the in-sample F-test is that it does not test for a post estimation sample predictive effect, so it is less credible as a true test of GC. The out-of-sample predictive test proposed herein is as follows.

First, the equation for y_t in (1) is rewritten as,

$$Y = Z\Phi + \varepsilon, \tag{2}$$

where Y is a $(T-p)\times 1$ vector containing $y_{p+1:T}$, ε is a $(T-p)\times 1$ vector $[u_{p+1:T}\ v_{p+1:T}]'$, Z is a $(T-p)\times(2p+1)$ matrix $[\mathbf{1}\ Y_{t-1}\ ...\ Y_{t-p}\ X_{t-1}\ ...\ X_{t-p}]$, with $\mathbf{1}$ a $(T-p)\times 1$ vector of 1s, Y_{t-j} a $(T-p)\times 1$ vector containing $y_{p-j+1:T-j}$ and X_{t-j} a $(T-p)\times 1$ vector containing $x_{p-j+1:T-j}$, and $\Phi = [\alpha_0\ \alpha_1\ ...\ \alpha_p\ \phi_1\ ...\ \phi_p]'$. To test whether y_t Granger causes x_t we rewrite the equation for x_t in (1) similar to (2), and test $H_0 : \beta_1 = ... = \beta_p = 0$.

For k-fold cross validation (with $k=1$ for LOO-CV), for each value of τ from $p+1$ to $T-k+1$, omit rows τ to $\tau+k+p-1$ from Y and Z to construct $Y_{-\tau}$ and $Z_{-\tau}$ such that they contain rows $p+1{:}\tau-1, \tau+k+p{:}T$ of Y and Z respectively. Adopting the standard Normal-Inverted Gamma prior (or Normal-Inverted Wishart for more general covariance matrix specifications) for the parameters of a model in the form of (2), the conditional posterior distributions are analytically tractable and well known: $\Phi|V \sim N(\bar{\Phi}, \Omega)$, $V|\Phi \sim IG(\nu/2, \delta/2)$, $\bar{\Phi} = \Omega[V^{-1}Z'_{-\tau}Y_{-\tau} + V_0^{-1}\Phi_0]$, $\Omega = [Z'_{-\tau}V^{-1}Z_{-\tau} + V_0^{-1}]^{-1}$, $\nu = \nu_0 + n$, $\delta = \delta_0 + (Y_{-\tau} - Z_{-\tau}\Phi)'(Y_{-\tau} - Z_{-\tau}\Phi)$, with Φ_0, V_0, ν_0, δ_0 prior parameters. The conditional posterior predictive distribution for out-of-sample predictions of Y is given by $\tilde{y}_\tau|\Phi, V \sim N(Z_{-\tau}\bar{\Phi}, \Omega)$, where $\tilde{y}_\tau$ is the predicted value for y_τ, so an MCMC sample is readily obtained from the Gibbs algorithm (Koop et al., 2007).

Using both k-fold cross validation and Gibbs sampling leads to the following algorithm.

Algorithm 1

1. For $\tau = 1+p : T$ and arbitrary starting value $V^{(0)}$:
    - (a) For $i = 1 : M+b$, generate draws
        - i. $\Phi^{(i)}|V^{(i-1)} \sim N(\bar{\Phi}, \Omega)$, with $V = V^{(i-1)}$ in $\bar{\Phi}$ and Ω,
        - ii. $V^{(i)}|\Phi^{(i)} \sim IG(\nu/2, \delta^{(i)}/2)$ with $\Phi = \Phi^{(i)}$ in δ,
        - iii. $\tilde{y}^{(i)}_{\tau:\tau+k-1}|Z_{-\tau}, \Phi^{(i)}, V^{(i)} \sim N(Z_{-\tau}\Phi^{(i)}, V^{(i)})$.
    - (b) Omit b burn-in draws, returning M post burn-in draws from the posterior predictive distribution for $y_{\tau:\tau+k-1}$.

2. Concatenate the $y_{\tau:\tau+k-1}$ draws to produce an $M \times T$ matrix,

$$\tilde{Y} = \begin{bmatrix} \tilde{y}_{p+1,1} & \tilde{y}_{p+2,1} & \cdots & \tilde{y}_{T,1} \\ \tilde{y}_{p+1,2} & \tilde{y}_{p+2,2} & \cdots & \tilde{y}_{T,2} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{y}_{p+1,M} & \tilde{y}_{p+2,M} & \cdots & \tilde{y}_{T,M} \end{bmatrix}. \qquad (3)$$

Applying Algorithm 1 to (2) with Z as defined above produces an ensemble of M realizations, $\tilde{Y}_U$, of the out-of-sample posterior predictive process for $y_{1:T}$ that allows for GC from x_t to y_t. Imposing the restrictions in the null hypothesis, equation (2) becomes,

$$Y = W\alpha + \omega, \tag{4}$$

where $W = [\mathbf{1}\ Y_{t-1}\ ...\ Y_{t-p}]$, $\alpha = [\alpha_0\ \alpha_1\ ...\ \alpha_p]'$, and $\omega = [u_{p+1:T}]'$. Defining $W_{-\tau}$ by omitting rows τ to $\tau + k + p - 1$ from W, allows application of Algorithm 1 to equation (4). This produces an $M \times T$ matrix $\tilde{Y}_R$, with the same structure as $\tilde{Y}_U$, but generated under the assumption of no GC from x_t to y_t.

Since the precision of parameter estimates decreases as the sample size is reduced, a logical approach is to set k as small as feasible. This will ensure that the precision of the model parameter estimates is close to that for the entire sample. This leads to $k = 1$, or LOO-CV, as the optimal choice of k provided the computational costs are not too great. However, one potential drawback of LOO-CV is that, if the null hypothesis is true, then consistent estimates of the parameters in the null hypothesis converge to zero asymptotically, so as $T \to \infty$, $\tilde{Y}_U \to \tilde{Y}_R$, suggesting that values of $k > 1$ may lead to more discriminative power for testing. This issue is explored in greater detail in Zhang and Yang (2015). As is demonstrated by the simulation results in Section 3.3, while LOO-CV is unlikely to be particularly computationally burdensome in standard applications, for reasonable sample sizes values of $k > 1$ can be selected without much loss of statistical power and with greater computational efficiency, so exploration over different values of k is a viable strategy.

For example, $\tilde{Y}_R$ for 120 observations generated from a simple AR(1) data generating process (DGP) given by (1) with $\alpha(B) = 1 - \alpha B \gamma(B) = 0, \phi(B) = 0, \beta(B) = 1, V = I_2$ is illustrated in Figure 1. In the Figure, the observed values for the dependent variable are represented by the black line, and the gray area is a plot of the prediction matrix, $\tilde{Y}_u$. Each grey line is a realization of the posterior predictive process, giving $M = 10^4$ draws from the predictive distribution for each $y_t, t = 1 + p : T$.

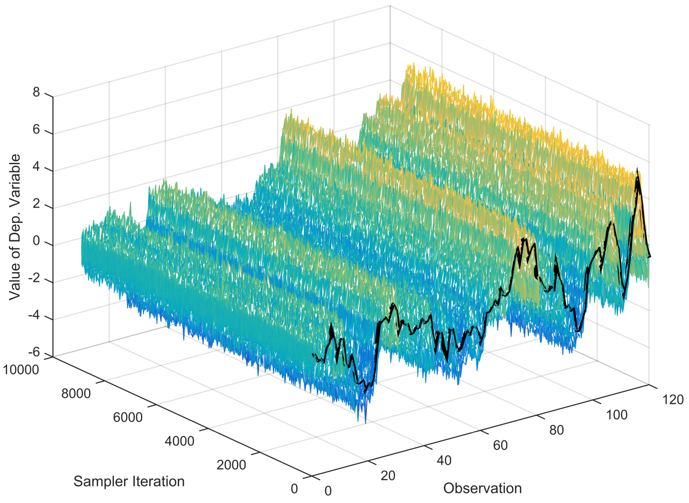


Figure 1: Predictive Ensemble Matrix, $\tilde{Y}_U$.

To determine the evidence against GC, the two predictive ensembles, $\tilde{Y}_U$ and $\tilde{Y}_R$, are compared by applying a loss function to obtain a statistical measure of average accuracy for each predictive realization, $\tilde{y}_{1:T,i}$, relative to the actual data $y_{1:T}$. The entire ensemble across the MCMC realizations allows computation of the exact posterior distribution of this statistic, which can then be used to test whether the additional information in x_t, improves predictive performance when predicting y_t.

The L2 norm distance is a standard choice of loss function for both estimation and prediction, leading to the root mean square prediction error measure (RMSPE), with the square root taken to scale the measure to match the predicted variable. L1 loss is also examined, which leads to the robust mean absolute error (MAE) measure of predictive performance.

When applied to a predictive ensemble $\tilde{Y}$, L2 loss can be expressed as,

$$RMSPE^{(i)} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_{t,i} - \tilde{y}_{t,i})^2}, \qquad i = (1, \ldots, M). \tag{5}$$

Applying (5) to $\tilde{Y}_U$ and $\tilde{Y}_R$ gives two $M \times 1$ vectors of draws from the posterior distributions for the RMSPE measures, $d_U = [RMSPE_U^{(1)}, ..., RMSPE_U^{(M)}]'$, and $d_R = [RMSPE_R^{(1)}, ..., RMSPE_R^{(M)}]'$. If the null hypothesis of no GC is false, these distributions will differ in location and possibly scale. If the null hypothesis is true, the distributions will have similar location, though the variance may differ due to the additional noise from nonzero parameter estimates of the extraneous parameters in the unrestricted model.

There are a number of potential ways to test if these two distributions differ. Since interest is in a null hypothesis of no GC, a test based on a comparison of means is a natural starting point. For the posterior means of $RMSPE_U$ and $RMSPE_R$, $\mu_U = \int zp(d_U = z|y,x)dz$ and $\mu_R = \int zp(d_R = z|y,x)dz$, the no GC hypothesis can be examined by testing $H_0 : \delta = \mu_R - \mu_U \leq 0$ vs. $H_0 : \delta > 0$.

To facilitate this comparison, the odds against the null hypothesis are calculated using an objective posterior odds ratio (Mills, 2018). This testing procedure does not suffer from the Jeffreys-Lindley-Bartlett paradox and allows the use of the same priors employed for posterior inference, including uninformative priors, so that scientific objectivity can be maintained. The outcome of the test is determined by the evidence from the data and any background information incorporated in the likelihood and prior. With a relatively uninformative prior, the prior has little to no influence on the test result.

Minimizing expected loss leads to the decision rule: reject H_0 if the posterior odds $O \geq L(H_1|H_0)/L(H_0|H_1) = c_0/c_1$, where $L(H_1|H_0) = c_0$ is the loss associated with choosing H_1 when H_0 is true (type I error), and $L(H_0|H_1) = c_1$ is the loss from choosing the null when the alternative is true (type II error). The posterior odds ratio is,

$$O = \frac{p(\delta > 0|x,y)}{p(\delta \leq 0|x,y)} \tag{6}$$

Critical odds ratio values, c_0/c_1, that approximately match 10%, 5% and 1% significance levels are 4:1, 7:1 and 30:1 respectively (Mills, 2018).

The testing procedure can be implemented by computing $p(\delta > 0|x,y))$ and $p(\delta \leq 0|x,y)$ from the psuedo-sample of M draws from the posterior for δ to obtain $p(\delta > 0|x,y))$ and $p(\delta \leq 0|x,y)$, then computing the posterior odds given by equation (6). This circumvents problems due to analytical intractability. The law of large numbers assures that the expected value of any function of the MCMC sample converges to its true value, i.e. for a sample of M draws for z, as $M \to \infty$,

$$\frac{1}{M}\sum_{i=1}^{M} f(z^{(i)}) \to E(f(z)), \tag{7}$$

where $z^{(i)}$ is the ith pseudo-sample draw. The accuracy of the simulated posterior density can be increased by increasing the pseudo-sample size, M. Chen (2005a) provides a detailed review of methods for computing posterior probabilities from an MCMC sample.

The distributional comparison of predictive performance can be extended to include higher moments since one can envision situations in which the restricted and unrestricted models produce predictions where $E(\delta|x, y) \approx 0$, but in the presence of more parameter uncertainty, have different variance or skewness for prediction errors. The derivation of the posterior odds from decision theoretic considerations allows for such an extension by modifying the test loss function. Setting $L(H_1|H_0)/L(H_0|H_1) = c_0\bar{\sigma}_U^2/c_1\bar{\sigma}_R^2$, weights the test decision loss function so that the loss associated with type I and II errors are weighted by the posterior variance of the d_U and d_R distributions, $\bar{\sigma}_R^2$ and $\bar{\sigma}_U^2$. Third and higher moments can be incorporated into the loss function in a similar fashion if desired.

Dividing both sides of the decision rule by this ratio leads to an augmented posterior odds ratio for evaluating the mean predictive error,

$$AO = \frac{\bar{\sigma}_R^2 p(\delta > 0|x, y)}{\bar{\sigma}_U^2 p(\delta \leq 0|x, y)} \tag{8}$$

This can lead to improved test performance in situations in which the mean prediction error is similar for both restricted and unrestricted models, but the variance of predictions for the unrestricted model is greater due to the inclusion of extraneous nonzero parameter estimates for parameters that are actually zero when the null hypothesis is true. The lower mean predictive variance for the restricted model then reduces the probability of a type I error. Similarly if either posterior density is skewed, the third moment can be incorporated into the loss function providing a further modified odds ratio. This can be important in forecasting situations for example, where avoidance of occasional large forecast errors may be more important than slightly reduced average forecast errors overall.

3 Monte Carlo Study

In this section, the small sample performance of the new GC testing procedure is evaluated in comparison to the in-sample F-test (and Wald test when heteroskedasticity is present), and the best performing out-of-sample test currently available in the literature, the Ashley and Tsang (2014) AT_{75} test. It is important to note that there are a number of alternative in-sample tests that have been proposed in the literature including, for example, nonparametric, frequency domain and quantile approaches (Candelon and Tokpavi, 2016, Liu and Moneaar, 2016). However, since the purpose of this paper is to propose and evaluate a new out-of-sample test in comparison to the best previously available out-of-sample test, for the sake of brevity these alternative in-sample procedures are not considered herein.

Empirical rejection rates are computed for each of the tests, with steadily increasing signal-to-noise ratio and the empirical test size fixed at 5% to allow power comparisons. The in-sample F-test is,

$$F = \frac{(RSS_R - RSS_U)/p}{RSS_U/(T-p-1)} \tag{9}$$

where $RSS_U = (Y - Z\bar{\Phi})'(Y - Z\bar{\Phi})$ and $RSS_R = (Y - W\bar{\alpha})'(Y - W\bar{\alpha})$. The Wald chi-square test is a commonly employed alternative to the in-sample F test as it is robust to potential heteroskedasticity. The Wald test was therefore included in the simulations evaluating a heteroskedastic data generating process. The Wald test is given by,

$$W = (R\Phi)'[R'V_{\Phi}R']^{-1}R\Phi \tag{10}$$

where V_{Φ} is the covariance matrix for $\hat{\Phi}$, and R is a $(T-p) \times (2p+1)$ matrix $[\mathbf{0}\ I_p\ \mathbf{0_p}]$ with I_p and 0_p a $p \times p$ identity matrix and matrix of zeros respectively.

The out-of-sample pseudo-F-statistic developed by Ashley and Tsang (2014) is computed by splitting the sample into two sub-samples at a partition point, τ, then using estimates from the sample $1:\tau$ to predict out-of-sample values $\tau+1:T$, and vice versa. While the choice of τ is arbitrary, Ashley and Tsang provide simulations that indicate an optimal choice of τ at the 75th percentile of the sample for empirical applications. We label this test statistic AT_{75}, which is given by,

$$AT_{75} \equiv \frac{(RSS_R^{\tau} - RSS_U^{\tau})/p}{RSS_U^{\tau}/(T-p-1)}, \tag{11}$$

where, $RSS_U^\tau = (Y_\tau - Z_\tau\bar{\Phi}_{-\tau})'(Y_\tau - Z_\tau\bar{\Phi}_{-\tau}) + (Y_{-\tau} - Z_{-\tau}\bar{\Phi}_\tau)'(Y_{-\tau} - Z_{-\tau}\bar{\Phi}_\tau)$, $RSS_R^\tau = (Y_\tau - W_\tau\bar{\alpha}_{-\tau})'(Y_\tau - W_\tau\bar{\alpha}_{-\tau}) + (Y_{-\tau} - W_{-\tau}\bar{\alpha}_\tau)'(Y_{-\tau} - W_{-\tau}\bar{\alpha}_\tau)$, with the subscript τ indicating the sample from 1:τ, and $-\tau$ indicating the sample $\tau + 1$:T. The distribution of the AT_{75} statistic is unknown; it is no longer F-distributed because negative F values occur when $RSS_R^\tau < RSS_U^\tau$ which cannot happen with the in-sample F-test, so either asymptotic distributional assumptions or bootstrapping is required to perform testing.

Ashley and Tsang employ the wild bootstrap (to address potential heteroskedasticity), however a drawback of the bootstrap in this setting is that an assumption of iid or exchangeable errors is required unless a moving block bootstrap is employed. The moving block bootstrap is difficult to implement with cross validation because whole blocks must be eliminated whenever one of the $k + p$ observations is contained in the block (k observations for out-of-sample prediction and p lags that contain one or more of those k observations). An advantage of MCMC methods here is that the ordering of the residuals is preserved, so inherent dependence remains and only the $k+p$ observations need to be omitted.

To evaluate the small sample performance of these procedures, each of the Monte Carlo simulations are carried out for sample sizes $T = 30, 60, 100$. For AO computation, an uninformative prior and $10,000$ iterations of the MCMC algorithm was used throughout sections 3 and 4.

The simulations have also been performed for larger samples, $T \geq 500$, and the results are consistent, though as expected all testing procedures converge to a power of one as T increases.

3.1 Two variable VAR processes

Consider the data generating process (DGP),

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \alpha_t & \phi \\ 0 & \gamma_t \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \quad t = 1, \ldots, T, \tag{12}$$

where both u_t and v_t are iid $N(0, 1)$. This DGP, from Chen (2005b), is examined in three scenarios. The first scenario is with time-invariant parameters, $\alpha_t = 0.3\ \forall\ t$ and $\gamma_t = 0.5\ \forall\ t$ (case 1). The second scenario has parameter instability from a structural break, $\alpha_t = \gamma_t = 0.2, t = 1, \ldots, \frac{T}{2}$ and $\alpha_t = \gamma_t = 0.8, t = \frac{T}{2} + 1, \ldots, T$ (case 2). As in Chen (2005b), different partitions were also considered for the structural break with consistent results. The third case extends case 1 to include heteroskedastic errors in the generation of y_t (case 3). This heteroskedasticity is in the form,

$$\mu_t = (\eta t)\epsilon_t, \quad \epsilon_t \sim N(0,1), \tag{13}$$

where $\eta = .01$ and $200 \leq t \leq 300$ leading to significant increase in variance of μ_t over the largest simulated data sets ($N = 100$).

Since the nominal size of the tests can be distorted if underlying test assumptions do not hold, particularly for modest sample sizes, the empirical size of all tests was fixed at 5% by simulation of critical values. To allow comparison of power with the odds ratio test (for which the nominal size is not fixed since it minimizes a linear combination of the type I and II errors), critical odds ratio values for *AO* were also computed to obtain an approximate 5% rejection rate when the null hypothesis is true. Table 1 shows the empirical size of each test by case and sample size. The clustering of the values around 5% indicates that the empirical rejection rates are comparable.

Table 1: Empirical Size (Nominal Size 5%)

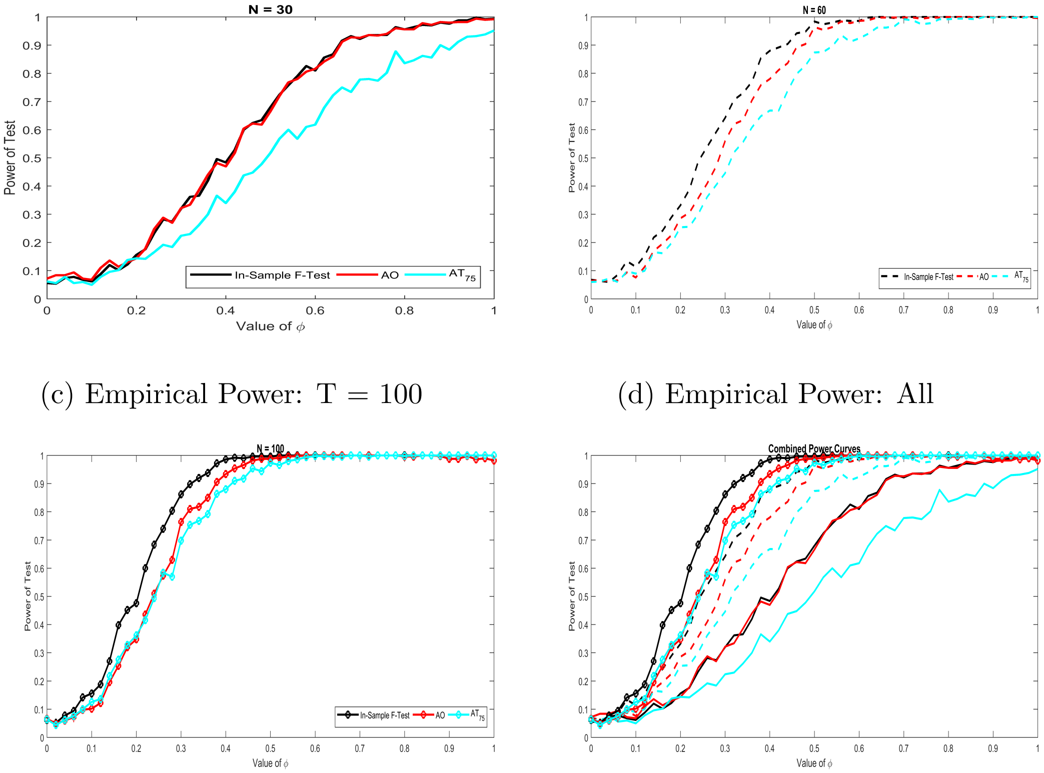
	Case 1			Case 2			Case 3		
Sample Size	30	60	100	30	60	100	30	60	100
F	0.048	0.050	0.050	0.052	0.051	0.048	0.048	0.053	0.052
Wald	-	-	-	-	-	-	0.052	0.049	0.049
AT_{75}	0.051	0.053	0.047	0.050	0.051	0.044	0.052	0.048	0.051
AO	0.050	0.057	0.043	0.057	0.053	0.050	0.040	0.050	0.048

Table 2: Empirical Power

	Case 1			Case 2			Case 3		
Sample Size	30	60	100	30	60	100	30	60	100
F	0.320	0.642	0.862	0.344	0.580	0.876	0.097	0.215	0.396
Wald	-	-	-	-	-	-	0.106	0.224	0.406
AT_{75}	0.224	0.446	0.698	0.170	0.362	0.520	0.123	0.161	0.180
AO	0.322	0.560	0.764	0.342	0.566	0.810	0.105	0.203	0.339

Table 2 shows the empirical rejection rates over 10^3 iterations when $\phi = 0.30$. Figures 2 and 3 present power curves for case 1 and 2. For example, Figures 2a to 2c provide power curves for each sample size examined by increasing the value of ϕ in increments of 0.02 over the interval of $[0, 1]$, with the empirical size fixed at 5% by selecting the critical value from the simulations when $\phi = 0$, and 500 replications performed at each increment. Figure 2d shows all three sample size power curves together. As expected, the increase in sample size shifts the curve up as parameter estimates become more precise and the posterior predictive distribution exhibits less variance. Figures 2d and 3d show that F and AO have roughly double the power of AT_{75} in terms of sample size required to attain a particular empirical rejection rate.

Figure 2: Case 1 Empirical Rejection Rates



(a) Empirical Power: T = 30

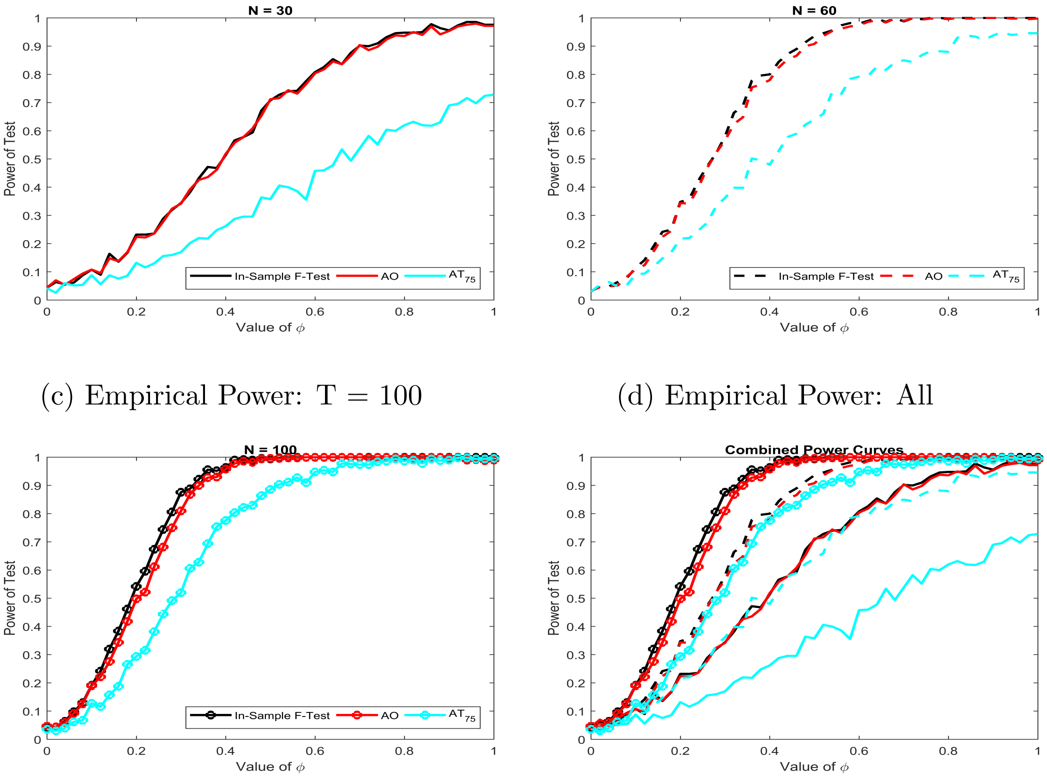
(b) Empirical Power: T = 60

(c) Empirical Power: T = 100

(d) Empirical Power: All

Across all three sample sizes and cases, the proposed test procedure exhibits greater statistical power than AT_{75}, with power close to that of the in-sample F-test. The results also indicate that structural breaks facilitate an increase in power for the out-of-sample tests.

Figure 3: Case 2 Empirical Rejection Rates



(a) Empirical Power: T = 30

(b) Empirical Power: T = 60

(c) Empirical Power: T = 100

(d) Empirical Power: All

3.2 Three variable VAR

To examine the effects of additional variables in the VAR process on test performance, an additional variable, z_t, was included in the VAR DGP,

$$\begin{bmatrix} y_t \\ x_t \\ z_t \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.8 \end{bmatrix} + \begin{bmatrix} 0.20 & \beta_x & \beta_z \\ 0.00 & 0.25 & 0.00 \\ 0.00 & 0.00 & 0.45 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t^y \\ \epsilon_t^x \\ \epsilon_t^z \end{bmatrix}, \begin{bmatrix} \epsilon_t^y \\ \epsilon_t^x \\ \epsilon_t^z \end{bmatrix} \sim MVN(\mathbf{0}, I_3). \tag{14}$$

Table 3: Empirical Size (Nominal Size 5%): Case 4

	$X \not\to Y$			$Z \not\to Y$		
Sample Size	30	60	100	30	60	100
F	0.050	0.050	0.048	0.052	0.054	0.051
AT_{75}	0.040	0.059	0.046	0.055	0.049	0.051
AO	0.052	0.064	0.066	0.051	0.053	0.062

Table 4: Empirical Power: Case 4

	$X \not\to Y$			$Z \not\to Y$		
Sample Size	30	60	100	30	60	100
F	0.306	0.623	0.846	0.323	0.668	0.896
AT_{75}	0.189	0.435	0.615	0.213	0.417	0.637
AO	0.296	0.577	0.753	0.299	0.612	0.817

Table 3 presents empirical size results from simulations designed to select an approximate 5% critical value for each test in this case. Table 4 presents the empirical rejection rates as T increases. The results are comparable to the two variable VAR specification with AO consistently outperforming AT_{75} by a substantial margin.

3.3 Misspecified model

In practice, the underlying DGP is unknown, so a model selection process must be undertaken prior to testing. This selection process typically involves choice of lag length, p, for the VAR model. The standard approach to lag length selection is to use model selection criteria, with the Akaike Information Criterion (AIC) and the Schwarz Criterion (BIC) being the most popular and the best performers in MC studies (Mills and Prasad, 1992; Cornwall and Mills, 2017). An alternative approach, explored in Cornwall and Mills (2017), is to use the LOO-CV out-of-sample prediction algorithm herein to compare predictive performance of VARs with different lag lengths, selecting the lag lengths with the best predictive performance. This alternative predictive lag length selection procedure is thus very similar to the GC testing procedure in Algorithm 1. Since none of these procedures is guaranteed to choose the best possible representation of the underlying DGP (and a VAR model is, in general, not the true DGP for actual data), some evidence on test performance when lag length is misspecified is provided below from an MC study using a different estimation and testing model to the true DGP. The DGP for this case is,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \varepsilon_t, \tag{15}$$

whereas the estimation and testing model is,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \varepsilon_t. \tag{16}$$

Table 5 provides the empirical rejection rates when the null hypothesis of no GC is true (Size) and false (Power). The results are again consistent with the previous results, with *AO* performance close to that of the in-sample *F* test, outperforming AT_{75} for all sample sizes. These results also provide evidence on the effect of increasing the number of observations omitted from estimation to allow out-of-sample prediction, *k*, or increasing the model lag length, *p*. Either *k* or *p* increasing leads to a larger number of observations being omitted from estimation (since $k + p$ observations must be omitted to prevent the information contained in the *k* observations from contaminating the estimation step through lagged values). As expected, as $k + p$ increases the power declines, though not substantially: for $T = 100$, with one lag of *x* and *y* (Table 2, case 2), the empirical rejection rate for *AO* is 0.810, whereas with an additional two lags of *x* and model misspecification, the rejection rate declines to 0.767 (Table 5).

Table 5: Misspecified Model: Size and Power

	Size			Power		
Sample Size	30	60	100	30	60	100
F	0.052	0.040	0.053	0.327	0.704	0.876
Wald	0.046	0.050	0.051	0.245	0.505	0.741
AT_{75}	0.059	0.056	0.047	0.202	0.462	0.589
AO	0.057	0.047	0.068	0.259	0.624	0.767

In addition to the results presented above, a number of other DGPs were explored, including those in Mills and Prasad (1992), McCracken (2007) and Ashley and Tsang (2014). The results were consistent across all DGPs considered, with the new predictive test consistently outperforming AT_{75}, showing minimal loss of power when compared to the in-sample *F*-test.

4 An Empirical Application

Granger causality testing has seen extensive empirical use in the context of the Phillips curve, evaluating evidence of a predictive relationship between inflation and unemployment (Atkeson and Ohanian, 2001; Clark and McCracken, 2006; Stock and Watson, 2007; Granger and Jeon, 2011). To evaluate this evidence using the proposed test procedure, data for the following variables were obtained: U.S. Personal Consumer Expenditures (PCE), U.S. Personal Consumer Expenditures excluding food and energy (PCE-Core) (U.S. Bureau of Economic Analysis, 2018b), U.S. Consumer Price Index - All goods (CPI) (Organization for Economic Co-operation and Development, 2018), U.S. Gross Domestic Product (GDP) (U.S. Bureau of Economic Analysis, 2018a), and U.S. Unemployment Rate (URT) (U.S. Bureau of Labor Statistics, 2018). All data are quarterly from 1963 (Q1) to 2018 (Q1), and were seasonally adjusted by the source agency.

The order of integration of each variable was determined by examining the standard deviation and autocorrelation plots pre and post differencing, and the Augmented Dickey-Fuller test. The results indicated that the inflation measures were all integrated of order 2, $I(2)$, whereas GDP and the unemployment rate were both $I(1)$, matching findings in previous studies. Appropriate lag lengths were selected using both AIC and BIC. Table 6 provides GC test results for the full sample. In the table, $\not\rightarrow$ denotes the null hypothesis of no GC, e.g. $H_0 : URT \not\rightarrow CPI$ states the hypothesis that the unemployment rate does not Granger cause the inflation rate as measured by the second difference of the CPI. The in-sample F-test indicates there is sufficient evidence to reject the null hypothesis for two out of the three measures of inflation (CPI and PCE). In contrast, the posterior odds, AO are much less than 4:1 which approximately matches a 10% sig. level (Mills, 2018), so the AO test indicates little evidence to suggest there is a Granger causal relationship for both L1 and L2 loss. As a benchmark, testing $H_0 : URT \not\rightarrow GDP$ and $H_0 : GDP \not\rightarrow URT$, which one would expect to reject, indicates strong to decisive evidence against the null hypothesis from both the AO (with odds of approx. 30:1 matching a 1% significance level)and the F test. For both the in-sample F-test and the AO test with L1 and L2 loss, the results remain unchanged if first differences of the inflation measures are employed.

Table 6: GC Test Results: Full Sample

Test	In-Sample F	P-Value	AO L1	AO L2
H_0 : URT $\not\to$ CPI	7.27	0.000	1.07	1.16
H_0 : URT $\not\to$ PCE	4.15	0.007	1.08	1.11
H_0 : URT $\not\to$ PCEc	1.93	0.127	1.00	1.01
H_0 : URT $\not\to$ GDP	36.2	0.000	227.7	73.1
H_0 : GDP $\not\to$ URT	35.0	0.000	84.5	38.0

The sample was also split into two sections, pre and post 1984, based on estimates of the great moderation (Stock and Watson, 2007, Atkeson and Ohanian, 2001). Table 8 shows test results for the post 1984 sample. The in-sample F-test provides sufficient evidence to reject the null hypothesis for all three measures of inflation. On the other hand, the *AO* test for both L1 and L2 loss indicates no evidence of Granger causality between the unemployment rate and inflation. Pre 1984 tests can be found in Table 7, where the results are more consistent across the tests, with the in-sample F-test also failing to reject the null hypothesis of no Granger causality for some measures of inflation.

The results from the proposed testing method differ from previous studies in finding insufficient evidence in both pre and post 1984 samples to reject the null hypothesis of no Granger Causality. This supports the assertion by Stock and Watson (2007) that “it has become much more difficult for an inflation forecaster to provide value added beyond a univariate model.” Interestingly, if the sample is restricted to only the 1960s, the quintessential example in the literature, then the new testing procedure indicates substantial evidence of a Granger causal relationship.

Table 7: GC Test Results: Post 1984 Sample

Test	In-Sample F	P-Value	AO L1	AO L2
H_0 : URT $\not\to$ CPI	2.67	0.050	1.04	1.04
H_0 : URT $\not\to$ PCE	3.72	0.013	1.00	1.05
H_0 : URT $\not\to$ PCEc	2.63	0.053	0.99	0.99
H_0 : URT $\not\to$ GDP	10.5	0.000	2.07	2.00
H_0 : GDP $\not\to$ URT	14.8	0.000	5.50	3.99

Table 8: GC Test Results: Pre 1984 Sample

Test	In-Sample F	P-Value	AO L1	AO L2
H_0 : URT $\not\rightarrow$ CPI	7.43	0.000	1.68	1.60
H_0 : URT $\not\rightarrow$ PCE	2.53	0.063	1.04	1.04
H_0 : URT $\not\rightarrow$ PCEc	1.04	0.378	1.04	1.01
H_0 : URT $\not\rightarrow$ GDP	24.9	0.000	119.2	60.2
H_0 : GDP $\not\rightarrow$ URT	21.0	0.000	37.7	18.7

5 Conclusion

In this paper, a new out-of-sample Granger causality testing procedure is presented that combines cross validation techniques with MCMC posterior simulation methods. A Monte Carlo study comparing empirical rejection rates for the new testing procedure with the in-sample F-test and the Ashley and Tsang (2014) 75th percentile pseudo-F-statistic indicates that the new test provides substantial improvement in statistical power over existing out-of-sample tests. Moreover, this new out-of-sample test, under conditions that are ideal for an in-sample test, produces power similar to the in-sample F-test. These simulation results suggest that out-of-sample predictive inference in a cross-validation framework can provide robust testing for Granger Causality even with modest sample sizes.

Out-of-sample tests avoid potential over-fitting and pretest estimator bias, and provide a more rigorous scientific procedure. Using the proposed methodology, the exact posterior predictive distribution for any test statistic can be obtained, allowing for out-of-sample predictive inference and testing in small samples. The test procedure is relatively easy to implement in practice (code is available at: https://github.com/tszanalytics/GrangerCausality) and has reasonably low computational cost. The procedure benefits from all the advantages of MCMC posterior simulation, and so can be readily extended to more complex nonlinear models.

The proposed procedure is applied to investigate the unemployment-inflation Phillips relationship in the U.S. using quarterly data from 1963:Q1 to 2018:Q1. Contrary to the findings from in-sample Granger causality testing, out-of-sample testing finds little evidence of a relationship.

References

Arlot, S., A. Celisse, et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys 4*, 40–79.

Ashley, R., C. W. Granger, and R. Schmalensee (1980). Advertising and aggregate consumption: an analysis of causality. *Econometrica*, 1149–1167.

Ashley, R. A. and K. P. Tsang (2014). Credible Granger-causality inference with modest sample lengths: a cross-sample validation approach. *Econometrics 2*(1), 72–91.

Atkeson, A. and L. E. Ohanian (2001). Are Phillips curves useful for forecasting inflation? *Quarterly Review 25*(1), 2.

Attanasio, A., A. Pasini, and U. Triacca (2012). A contribution to attribution of recent global warming by out-of-sample Granger causality analysis. *Atmospheric Science Letters 13*(1), 67–72.

Barnett, L., A. B. Barrett, and A. K. Seth (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters 103*(23), 238701.

Barnett, L., A. B. Barrett, and A. K. Seth (2017). Reply to Stokes and Purdon: A study of problems encountered in Granger causality analysis from a neuroscience perspective. *arXiv preprint arXiv:1708.08001*.

Barnett, L. and A. K. Seth (2015). Granger causality for state-space models. *Physical Review E 91*(4), 040101.

Chen, M.-H. (2005a). Bayesian computation: From posterior densities to Bayes factors, marginal likelihoods, and posterior model probabilities. In D. Dey and C. Rao (Eds.), *Bayesian Thinking*, Volume 25 of *Handbook of Statistics*, pp. 437 – 457. Elsevier.

Chen, S.-S. (2005b). A note on in-sample and out-of-sample tests for Granger causality. *Journal of Forecasting 24*(6), 453–464.

Clark, T. E. (2004). Can out-of-sample forecast comparisons help prevent overfitting? *Journal of Forecasting 23*(2), 115–139.

Clark, T. E. and M. W. McCracken (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics 105*(1), 85–110.

Clark, T. E. and M. W. McCracken (2005). The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics 124*(1), 1–31.

Clark, T. E. and M. W. McCracken (2006). The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence. *Journal of Money, Credit, and Banking 38*(5), 1127–1148.

Cornwall, G. and J. Mills (2017). Prediction based model selection criteria. manuscript.

Dhamala, M., G. Rangarajan, and M. Ding (2008). Estimating Granger causality from Fourier and wavelet transforms of time series data. *Physical Review Letters 100*(1), 018701.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Diks, C. and M. Wolski (2016). Nonlinear Granger causality: Guidelines for multivariate analysis. *Journal of Applied Econometrics 31*(7), 1333–1351.

Ghysels, E., J. B. Hill, and K. Motegi (2016). Testing for Granger causality with mixed frequency data. *Journal of Econometrics 192*(1), 207–230.

Granger, C. W. and Y. Jeon (2011). The evolution of the Phillips curve: a modern time series viewpoint. *Economica 78*(309), 51–66.

Hu, X., S. Hu, J. Zhang, W. Kong, and Y. Cao (2016). A fatal drawback of the widely used Granger causality in neuroscience. In *Information Science and Technology (ICIST), 2016 Sixth International Conference on*, pp. 61–65. IEEE.

Inoue, A. and L. Kilian (2005). In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews 23*(4), 371–402.

Koop, G., D. J. Poirier, and J. L. Tobias (2007). *Bayesian Econometric Methods*. Cambridge University Press.

Liao, W., J. Ding, D. Marinazzo, Q. Xu, Z. Wang, C. Yuan, Z. Zhang, G. Lu, and H. Chen (2011). Small-world directed networks in the human brain: multivariate Granger causality analysis of resting-state fMRI. *Neuroimage 54*(4), 2683–2694.

McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics 140*(2), 719–752.

Mills, J. (2018). Objective Bayesian Precise Hypothesis Testing. manuscript.

Mills, J. A. and H. Namavari (2017). Objective Bayesian ANOVA Testing. *University of Cincinnati*.

Mills, J. A. and K. Prasad (1992). A comparison of model selection criteria. *Econometric reviews 11*(2), 201–234.

Organization for Economic Co-operation and Development (2018). Consumer price index: Total all items for the United States.

Picard, R. R. and R. D. Cook (1984). Cross-validation of regression models. *Journal of the American Statistical Association 79*(387), 575–583.

Poirier, D. J. (1988). Causal relationships and replicability. *Journal of Econometrics 39*, 213–234.

Roebroeck, A., E. Formisano, and R. Goebel (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage 25*(1), 230–242.

Roebroeck, A., E. Formisano, and R. Goebel (2011). The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *Neuroimage 58*(2), 296–302.

Seth, A. (2007). Granger causality. *Scholarpedia 2*(7), 1667.

Sims, C. A. (1972). Money, income, and causality. *The American Economic Review 62*(4), 540–552.

Stock, J. H. and M. W. Watson (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking 39*(s1), 3–33.

Stokes, P. A. and P. L. Purdon (2017). A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proceedings of the National Academy of Sciences 114*(34), E7063–E7072.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, 29–35.

U.S. Bureau of Economic Analysis (2018a). Table 1.1.6. Real Gross Domestic Product, Chained Dollars.

U.S. Bureau of Economic Analysis (2018b). Table 2.4.4U Price Indexes for Personal Consumption Expenditures by Type of Product.

U.S. Bureau of Labor Statistics (2018). Civilian Unemployment Rate.

Yu, L., J. Li, L. Tang, and S. Wang (2015). Linear and nonlinear Granger causality investigation between carbon market and crude oil market: A multi-scale approach. *Energy Economics 51*, 300–311.

Zhang, D. D., H. F. Lee, C. Wang, B. Li, Q. Pei, J. Zhang, and Y. An (2011). The causality analysis of climate change and large-scale human crisis. *Proceedings of the National Academy of Sciences 108*(42), 17296–17301.

Zhang, Y. and Y. Yang (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics 187*(1), 95–112.

Acknowledgements

The authors gratefully acknowledge support from the Taft Fund, University of Cincinnati, and feedback from Rick Ashley and participants at the Midwest Econometrics Group Meetings. Any views expressed here are those of the authors and not necessarily those of the Bureau of Economic Analysis or U.S. Department of Commerce.