Exploring Granger Causality in Econometric Data: Model Selection, Significance Tests, and Penalized Regression

Sikata B. Sengupta
Department of Mathematics and Economics
Stanford University
sikata@stanford.edu

Abstract

Granger causality is a common way to look for relationships between variables in time series data. In a complex model, detection of Granger causality is intimately entangled with other issues like lag selection. These problems could be formulated as identifying a structural vector autoregressive (SVAR) model. This can in turn be reformulated as a regression problem with variable selection. Our exploration of Granger causality is motivated by questions regarding the causal relationship between consumption growth on GDP growth and vice-versa. Potential approaches here involve fitting VAR models of different orders using various information criteria, hypothesis testing, penalized regression with group LASSO, and with hierarchical group LASSO. We consider how each of these models fare on synthetic data, with features similar to U.S. econometric data. We will see that the cross-validation-based criterion, used for prediction, is too lax for variable selection. Finally, we explore variable selection in the hierarchical group LASSO approach with a procedure inspired by the knockoff filter. We find that even this procedure is not stringent enough for setting the right value of the penalty parameter, at which hierarchical group LASSO is capable of finding the right structure.

1 Introduction

Granger [6] formulated a notion of temporal causality, based on the idea that a cause occurs before its effect and that the knowledge of a cause improves the prediction of its effect. This formulation has seen many applications in the field of time series. Granger's notion was originally developed with econometric applications in mind. In this project, I will explore different ways of inferring Granger causality, staying within vector autoregressive models (VARs) [11].

A relationship that is often investigated in macroeconomics is the connection between consumption growth and GDP growth [13]. The standard National Accounts Identity: $Y = C + I + G + NX$, where Y, C, I, G, NX are output, consumption, investment, government spending, and net exports respectively, would suggest there is some relationship between consumption growth and GDP growth. However, characterizing whether consumption or GDP growth drives the other is a nontrivial task.

Traditional F-Test Granger Causality methods have been applied to this problem earlier, as in [7]. We will extend this analysis by studying how group lasso [14] and hierarchical vector autoregression (HVAR/HLag) [9, 10], a particular form of overlapping group lasso, perform in various lag selection and Granger causality settings, through synthetic data generated based upon consumption and GDP growth rates. We also test a knockoff-based approach [1] for variable selection with HVAR.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Background

2.1 Conventional approach to lag selection and Granger causality in VARs

A p-th order d-dimensional VAR for variables $y_t \in \mathbb{R}^d$ is defined as follows:

$$y_t = \nu + \sum_{i=1}^{p} \Phi_l y_{t-l} + u_t, \text{ where } u_t \sim \mathcal{N}(0, \Sigma).$$

In practice, the order p is decided by one of many model selection criteria that balances goodness of fit against model complexity.

If one assumes Σ to be proportional to the identity, one can rephrase the problem as regression via ordinary least squares (OLS):

$$(\hat{\beta}, \hat{\nu}) = \underset{\beta,\nu}{\operatorname{argmin}} ||\mathbf{y} - \boldsymbol{\nu} - \mathbf{X}\beta||_F^2$$

where

$$\mathbf{y} = \begin{pmatrix} y_{l+1}^T \\ y_{l+2}^T \\ \vdots \\ y_n^T \end{pmatrix}, \ \boldsymbol{\nu} = \begin{pmatrix} \nu^T \\ \nu^T \\ \vdots \\ \nu^T \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} y_l^T & y_{l-1}^T & \cdots & y_1^T \\ y_{l+1}^T & y_l^T & \cdots & y_2^T \\ \vdots & \vdots & \cdots & \vdots \\ y_{n-1}^T & y_{n-2}^T & \cdots & y_{n-l}^T \end{pmatrix}, \text{ and, } \beta = \begin{pmatrix} \Phi_1^T \\ \Phi_2^T \\ \vdots \\ \Phi_l^T \end{pmatrix}$$

and $||.||_F$ is the Frobenius norm for matrices: $||A||_F = \sqrt{\sum_{ik} |A_{ik}|^2}$.

To check if component $y_{jt'}$ affects y_{it} for $t > t'$, one considers two models: One with β corresponding to full Φ_l's and another with β_-, where $(\Phi_l)_{ij} = 0$ for all $l = 1, \ldots, p$. An F-test checking whether adding $(\Phi_l)_{ij}$'s improved the fit significantly is one way of detecting Granger causality.

2.2 Group lasso for Granger causality

Group lasso [14] applied to the problem is formulated as

$$(\hat{\beta}, \hat{\nu}) = \underset{\beta,\nu}{\operatorname{argmin}} ||\mathbf{y} - \boldsymbol{\nu} - \mathbf{X}\beta||_F^2 + \lambda \mathcal{P}_g(\beta)$$

where the penalty

$$\mathcal{P}_g(\beta) = \sum_{i=1}^{d} \sum_{j=1}^{d} ||((\Phi_1)_{ij}, (\Phi_2)_{ij} \ldots, (\Phi_p)_{ij})||_2 := \sum_{i=1}^{d} \sum_{j=1}^{d} ||((\Phi_{1:p})_{ij})||_2$$

encourages setting unimportant ij interactions to zero, across lags. We use the notation $(\Phi_{l:p})_{ij}$ to indicate the vector $((\Phi_l)_{ij}, (\Phi_{l+1})_{ij} \ldots, (\Phi_p)_{ij})$.

It is common to perform λ selection by cross-validation.

2.3 HVAR for combining lag selection and Granger causality detection

If, in addition, we want to perform lag selection we want to use overlapping group lasso penalizing group (ij) coefficients associated with higher lags. One way to achieve this is regression with an elementwise hierarchical penalty $\mathcal{P}_h(\beta)$ [9, 10]:

$$(\hat{\beta}, \hat{\nu}) = \underset{\beta,\nu}{\operatorname{argmin}} ||\mathbf{y} - \boldsymbol{\nu} - \mathbf{X}\beta||_F^2 + \lambda \mathcal{P}_h(\beta)$$

where

$$\mathcal{P}_h(\beta) = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{l=1}^{p} ||((\Phi_{l:p})_{ij})||_2.$$

Once more, λ is often fixed by cross-validation.

Having discussed existing methods, we propose two other exploratory approaches as well, in the next two sections.

3 HVAR penalty selection with knockoffs for λ selection

After the advent of lasso [12] it has been used both for regression shrinkage, namely, making coefficients smaller so as not to be too sensitive to noise, and therefore improve prediction, and for variable selection, since the procedure can set certain coefficients to zero. However, over time, it became clear that there is some conflict between optimal prediction and consistent variable selection in lasso [5, 8]. If lasso is optimized for prediction, as would be the case for cross-validation, the lasso typically picks true variables, along with many null-variables, coming with small estimated coefficients. These issues lead to the developments approaches like adaptive lasso [15], more suited towards variable selection.

More recently, the knockoff-filter [1, 3] has become a major tool for variable selection with false discovery rate (FDR) control. The idea for the knockoff filter is to generate 'knockoff features' $\tilde{X}$ along with true features X. For each feature one compares the 'strength' of the regression coefficients with that of the corresponding 'knockoff feature'. If the knockoff feature has larger absolute value of coefficients, it should indicate some probability of the feature being a null-feature, meaning the corresponding coefficient is zero. Dai and Barber suggested an adaptation of the knockoff filter for group sparsity [4]. Knockoff filters are not usually utilized for selection of penalty strength. That is taken care of by something like cross-validation. Only then variable selection is done by the knockoff filter for a particular setting of FDR [1].

We propose a procedure for the choice of λ for variable selection for our penalized regression problem:

1. Select fixed-X knockoff $\tilde{X}$ as described in [1],
2. Perform regression with an elementwise hierarchical penalty, with different λs,
3. Find the the minimum λ, λ_m where all knockoff coefficients have lower absolute value than the absolute value of the corresponding true feature coefficients,
4. Output which coefficients are non-zero at λ_m.

4 Bootstrapping for λ selection

A second simple alternative is to generate synthetic data where the ground truth is known. Fit penalized regression models with the the highest variable selection accuracy. Then run the penalized regression model on the original data.

More precisely:

1. Fit data to a structurally restricted VAR (SVAR) where some coefficient(s) has been set to zero,
2. Run penalized regression for different λs on synthetic data generated from the SVAR,
3. Find $\lambda = \lambda_o$ corresponding to the highest accuracy of calling null and non-null features in the generative SVAR,
4. Output which coefficients are non-zero at λ_0.

5 Experiments

For macroeconomic studies, one can use publicly available data from the Bureau of Economic Analysis (`https://www.bea.gov/`) and from the OECD database (`https://data.oecd.org/`). For early testing of my code, I have used pre-processed quarterly data on U.S. consumption and GDP growth during the period between 1990 and 2019 from a tutorial on conventional approaches to Granger causality [7]. First, we visualize the consumption and GDP data given and their corresponding growth rates in Fig. 1.

5.1 Conventional lag selection by information criteria depends on the criterion

We then fit a VAR(p) [11] on consumption and GDP growth. The order p was selected by searching up to a maximal lag of 5. The Akaike Information Criterion (AIC) [2] picks $p = 3$. Alternative

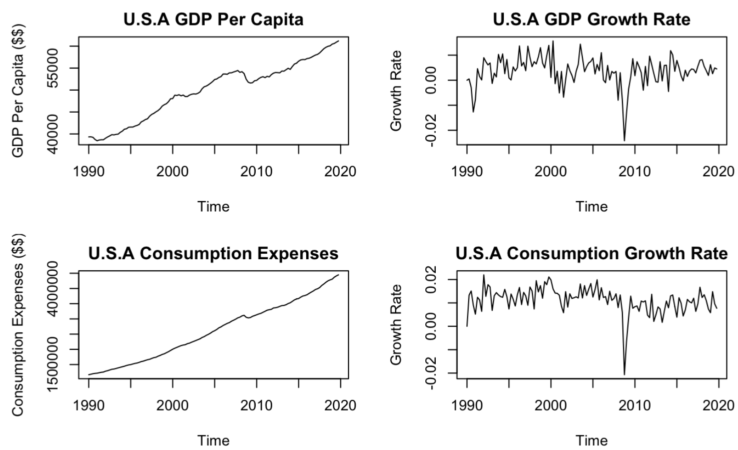

Figure 1: Plots of GDP, Consumption, and their growth rates every quarter from 1990 to 2019.

information criteria, like Bayesian information criterion (BIC) [2], suggest lower lag values like $p = 1$. However, since in the future, we are generally interested in looking at classes of models that automatically do lag selection and penalize higher order lag inclusions, we decided to go with $p = 3$.

5.2 Granger causality tests suggests marginal effect of GDP growth

Now we run two F-tests on our VAR(3) model for Granger causality : does consumption growth Granger cause GDP growth and vice-versa?

Causation Variable	p-value
Consumption Growth	0.4408
GDP Growth	0.05638

The first test gives us a p-value of 0.4408, and thus we fail to reject the null hypothesis that consumption growth does not Granger cause GDP growth. In the reverse test however, we get a p-value of 0.05638 which is right on the edge depending on whether you choose a significance level of 0.05 or higher. Though one may end up concluding that even in this case we fail to reject the null hypothesis that GDP does not Granger cause consumption growth, it is clear that there seems to be more significance of GDP's role in predicting consumption than vice versa.

5.3 Group lasso approach indicates small effects in both directions

While these hypothesis tests are indeed useful, they depend upon assumptions on the generative model, especially on the structure of noise. A more flexible approach is to treat this problem as a supervised learning problem of predicting the future out of various lagged features.

We apply penalized regression via the group LASSO approach– mainly focusing on predicting consumption growth out of GDP growth. Grouping involves all regression coefficients for influence of a variable for multiple lags. In Fig. 2, we plot the norm of our parameter vector β for both consumption and GDP growth against various different values of our penalty coefficient. We can see that at very high values (on the left hand side of the image) of penalty, the norm of the parameter growth vector gets activated first for consumption, and then for GDP growth. Now, we run cross-validation to find the optimal penalty strength and then choose the resulting fit based upon that optimal penalty ($4.85102762467957e - 07$). See Fig. 2. In the resulting fit, we find that group LASSO ends up putting smaller weight on farther-back lags of GDP-growth rate (though non-zero) but finds both lag-one values to be meaningful in predicting the consumption growth rate.

We also perform regression GDP growth in terms of both variables and find nonzero coefficients for consumption growth. In summary, if we take the group lasso results at the cross-validation optimized point, we will declare both variables Granger causing each other, in contrast to the results from the F-tests.

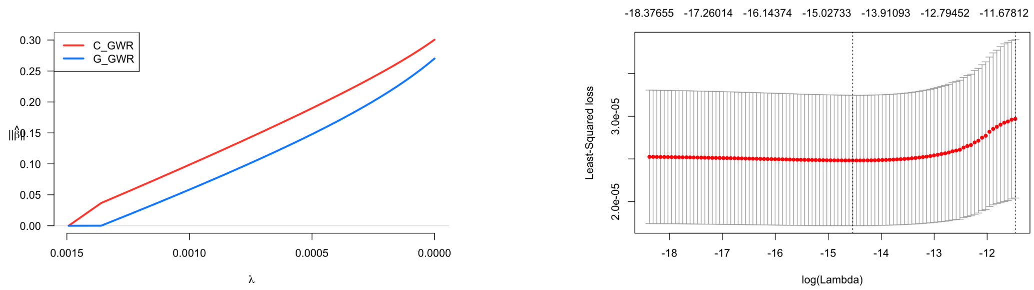

Figure 2: On the left is a plot of β norm vs λ values and on the right is a plot of the least squares loss vs λ (with the x-axis decreasing from left to right).

5.4 Tests on synthetic data

We have seen that our preliminary analysis suggests GDP growth might Granger cause consumption growth. It would be worthwhile to check the relative merit/demerits of either method, the hypothesis testing based one and the penalized regression based one in a case where the ground truth is known about the presence or the absence of Granger causality. This enables us to understand the strengths and weaknesses of each method before proceeding to consider alternative/more nuanced techniques.

Hence, we will now shift to running a few experiments on synthetic data to test how Granger causality F-tests and group LASSO perform in settings where we perfectly know the model that generated the data. In order that the synthetic data generation models stay close to real data, we will pick our generative models by fitting data with VAR models with different structures.

5.5 Simulations from estimated VAR(3) and sensitivity to covariance structure

First, we simulate 100 data points according to our estimated VAR(3) model above in two different ways. In one, we use the estimated covariance matrix of our residuals to be the noise matrix, and the other we use the default multivariate normal identity matrix for noise. We perform Granger causality tests on the effect of GDP growth on consumption after fitting a VAR(3) model on both simulated datasets. In the case where we use the covariance matrix of residuals, since those values are incredibly tiny, we get a p-value of 0.2515 and fail to reject the conclusion that GDP growth does not Granger cause consumption growth. However, when we use the default noise setting specified above, we get a p-value of 0.003494 in which case we reject the null hypothesis and conclude that GDP growth does Granger cause consumption growth. This experiment ultimately highlighted the fact that the Granger causality conclusions we draw are highly sensitive to the nature of the noise specification.

We then also try our synthetic dataset (that was created with the covariance of the residuals matrix) and see that it fits a model where it puts non-zero coefficients on all lagged values of all variables provided, enabling us to conclude in that GDP growth does Granger cause consumption growth in the group-LASSO case, and that group lasso is more robust to noise specification than the VAR-fitting Granger F-test approach.

Next, we try another type of synthetic data experiment. Here, we simulate synthetic data, according to two independent AR(3) processes obtained by fitting independently consumption and GDP growth. We then see whether both the Granger causality F-test and group LASSO approach conclude that, in this synthetic dataset, neither variable Granger causes the other.

In this case, after fitting a VAR(3) and applying our Granger F-test, we get a p-value of $.5976$, and thus fail to reject the null hypothesis that GDP growth does not Granger cause consumption growth. Our group LASSO approach ends up placing relative-small, but yet nonzero coefficient values on the lagged GDP values, which in the naive scheme of determining Granger causality, would end up making it conclude that GDP growth Granger causes consumption growth. This experiment ended up highlighting that depending on how we look to determine Granger causality in the group LASSO scheme (normally zero or non-zero coefficients), the approach is rather sensitive to rather small, but nonzero coefficients placed on independent series.

6 Simulations from two estimated independent AR(3)'s

We will perform many procedures on a synthetic dataset in which both variables have no influence on each other. To this end, we estimate two independent AR(3)'s, one from consumption growth data, and another from GDP growth data. Thus our ground truth has lag $p = 3$ and there is not Granger causality between the variables, we keep a maximum allowed lag of 5, and test which methods come close to detecting the true structure of the VAR.

6.0.1 Lag selection by information criteria

We use the same aforementioned selection criterion to determine the order p of the VAR we fit on our synthetic data. AIC selects $p = 4$ and BIC chooses $p = 1$, while the true order is $p = 3$.

6.0.2 F-tests detect lack of causal influence

We run two F-tests on our lag 3 VAR model on synthetic data.

Causation Variable	p-value (lag 1)	p-value (lag 3)	p-value (lag 4)
Cons. Growth	0.8956	0.6386	0.4844
GDP Growth	0.6978	0.4689	0.3266

F-tests correctly picks lack of causal influence. Even if we provide the wrong lag, for example those suggested information criteria, the p-values change but the lack of significance remains.

6.0.3 Group lasso does not rule out any causal interaction

We performed group lasso for lag 3. All 3 lags of both variables activated so, superficially both variables Granger cause each other. This lack of stringency in variable selection is consistent with our previous experience running group lasso on real data.

6.0.4 HVAR can perform well variable selection if the right λ is chosen

We perform HVAR with elementwise penalty for many values of λ with maximum lag 5. We score the results by the accuracy with which it detects null and non-null variables, just as in classification problems. Fig. 3 shows accuracy as a function of λ and also the ROC curve with λ as the parameter being scanned. We find $-\log_{10}\lambda \approx 1.3$ to be rather accurate for variable selection.

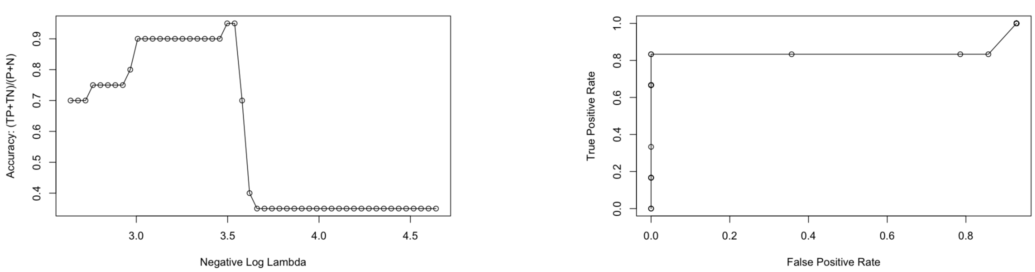

Figure 3: The left side is a plot of accuracy against values of λ and the right side plot is a plot of the ROC curve for the HVAR procedure.

6.0.5 Group lasso with knockoff procedure does not select a good λ

As proposed before, we want to use fixed-X knockoffs to guide the the choice of λ. Fig. 4 shows that around $-\log_{10}\lambda$ in the range of $1.5 - 2.5$ shows some knockoff features being picked up in preference to true features. However, Fig. 5 suggests that higher a higher value of λ, $-\log_1 0\lambda \approx 1$ has higher accuracy.

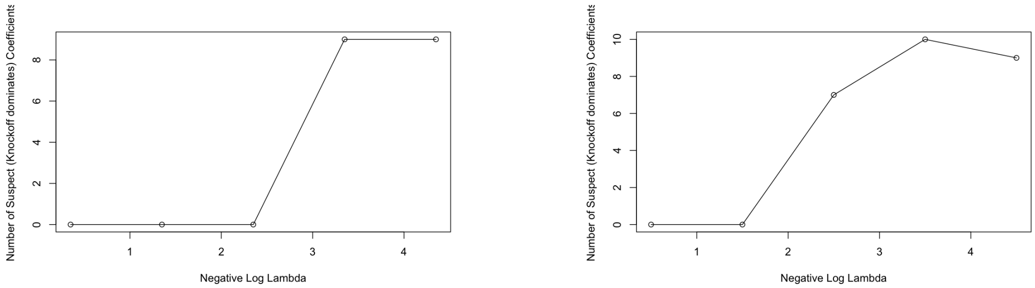

Figure 4: Both plots broadly show the relation between the number of suspect coefficients (knockoff value dominates actual value) and λ values for the knockoff procedure. The left side plot is for that of consumption growth, and the right is for that of GDP growth.

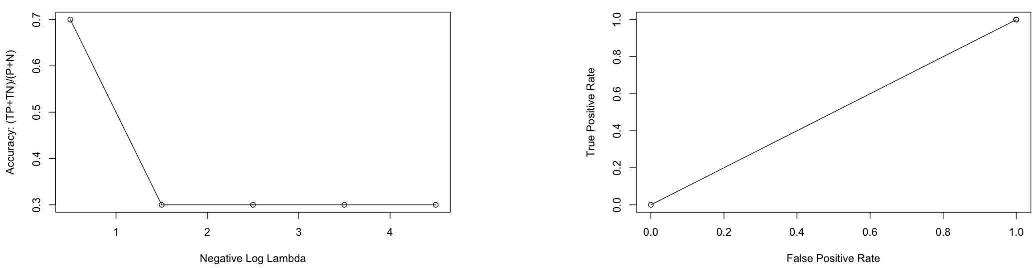

Figure 5: The left side is a plot of accuracy against values of λ and the right side plot is a plot of the ROC curve for the knockoff procedure.

6.0.6 HVAR with knockoff does not select a good λ either

Now we use fixed-X knockoffs to guide the the choice of λ for HVAR. Fig. 4 shows that around $-\log_{10}\lambda$ in the range of $1.5 - 2.5$ shows some knockoff features being picked up in preference to true features. However, Fig. 6 suggests that higher a higher value of λ, from strong activation of knockoffs, compared to the value of λ where accuracy is the highest (indicated by arrows). The accuracy profile and the ROC curve can be seen in Fig. 7.

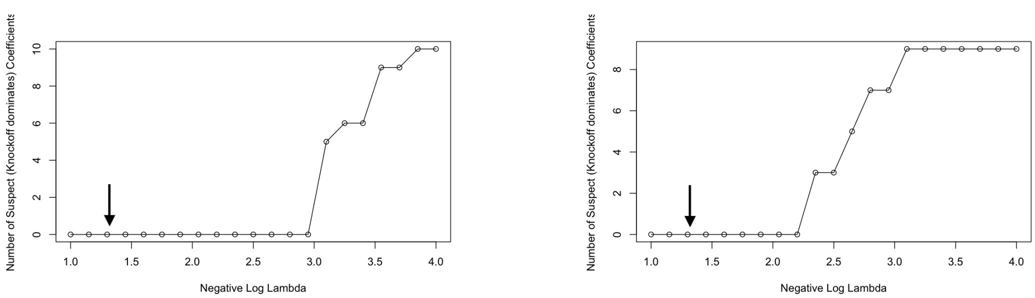

Figure 6: Both plots broadly show the relation between the number of suspect coefficients (knockoff value dominates actual value) and λ values for the procedure of HVAR plus knockoff. The left side plot is for that of consumption growth, and the right is for that of GDP growth. The arrow points to the value of 1.3 which corresponds to the value at which the accuracy plot peaked for this procedure.

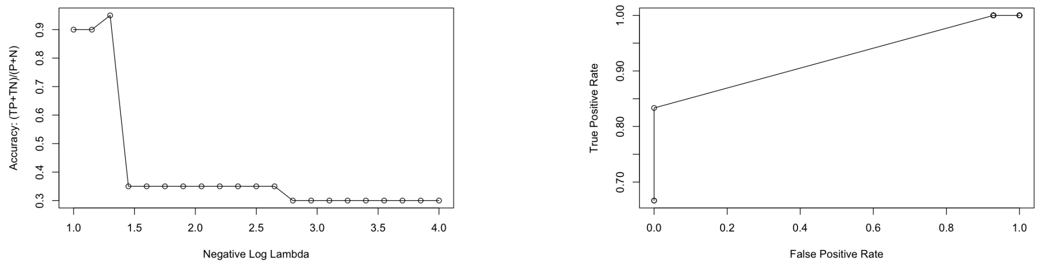

Figure 7: The left side is a plot of accuracy against values of λ and the right side plot is a plot of the ROC curve for the HVAR plus knockoff procedure.

6.1 Bootstrap to pick λ and fitting actual data

Finally, we use the optimal λ from investigation of HVAR accuracy, to the original dataset. We will call cgw, ggw the growth rates of consumption and of GDP, respectively.We explicitly show the fit

$$\begin{bmatrix} cgw_t \\ ggw_t \end{bmatrix} = \begin{bmatrix} 0.1872 & 0.1940 \\ 0.1798 & 0.1184 \end{bmatrix} \begin{bmatrix} cgw_{t-1} \\ ggw_{t-1} \end{bmatrix} + \begin{bmatrix} 0.0338 & 0.0000 \\ 0.0000 & 0.0527 \end{bmatrix} \begin{bmatrix} cgw_{t-2} \\ ggw_{t-2} \end{bmatrix} + \begin{bmatrix} 0.0380 & 0.0000 \\ 0.0000 & 0.0000 \end{bmatrix} \begin{bmatrix} cgw_{t-3} \\ ggw_{t-3} \end{bmatrix} + \begin{bmatrix} 0.0010 & 0.0000 \\ 0.0000 & 0.0000 \end{bmatrix} \begin{bmatrix} cgw_{t-4} \\ ggw_{t-4} \end{bmatrix} + \begin{bmatrix} 0.0002 & 0.0000 \\ 0.0000 & 0.0000 \end{bmatrix} \begin{bmatrix} cgw_{t-5} \\ ggw_{t-5} \end{bmatrix} + \begin{bmatrix} 0.0076 \\ 0.0014 \end{bmatrix} + \begin{bmatrix} \text{noise} \\ \text{noise} \end{bmatrix}. \quad (1)$$

From this fit, one might imagine the $p = 1$ lag selected by BIC may have some justification. However, at that level, we are seeing causal effects in both directions.

7 Future Directions and Conclusion

The efficacy of HVAR [9] to perform variable selection with the appropriate λ calls for further investigation. It is possible that a variation of methods like adaptive lasso, that are geared toward variable selection can succeed in making HVAR into a very useful tool. In addition, even if our first effort to utilize knockoffs failed, developing a principled knockoff filter appropriate for time series data remains an interesting direction.

References

[1] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[2] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

[3] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[4] Ran Dai and Rina Barber. The knockoff filter for fdr control in group-sparse and multitask regression. In *International conference on machine learning*, pages 1851–1859. PMLR, 2016.

[5] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[6] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[7] Arimitra Maiti. Fun with ARMA, VAR, and Granger Causality. `https://towardsdatascience.com/fun-with-arma-var-and-granger-causality-6fdd29d8391c`, 2020. [Online; accessed 16-November-2021].

[8] Nicolai Meinshausen and Peter Bühlmann. Variable selection and high-dimensional graphs with the lasso. *Ann Stat*, 34:1436–1462, 2006.

[9] William Nicholson, David Matteson, and Jacob Bien. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094*, 2017.

[10] William B Nicholson, Ines Wilms, Jacob Bien, and David S Matteson. High dimensional forecasting via interpretable vector autoregression. *J. Mach. Learn. Res.*, 21:166–1, 2020.

[11] Robert H Shumway and David S Stoffer. *Time series: a data analysis approach using R*. Chapman and Hall/CRC, 2019.

[12] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[13] Yi Wen. Granger causality and equilibrium business cycle theory. *FRB of St. Louis Working Paper No*, 2005.

[14] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[15] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.