Quality checks of Granger causality libraries

in the bivariate and multivariate case

University of Passau
School of Business, Economics and Information Systems

Financial Data Analytics
Prof. Dr. Ralf Kellner

Type:	Seminar paper
Study program:	Business Administration, M.Sc.
Submitted by:	Florian Peschke Enrollment number: 105259 E-Mail: florian.peschke@uni-passau.de
Submitted:	February 11, 2022
Supervisor:	Jan König
In collaboration with:	Danilo Saft (NORD/LB)

Appended files/folders:	`CausalA`
	`empirical results`
	`bivariate.ipynb`
	`multivariate.ipynb`
	`main.r`

Contents

List of Tables

List of Figures

1 Introduction

Causality and correlation are two terms that are closely related, but differ in their meaning. If we consider a context with two random variables X and Y, the correlation between X and Y indicates the relationship between them, but it gives no indication of the direction of that relationship. That is, X can be used to explain Y and vice versa (Granger, 1988). It may even be a random relationship due to various factors not directly considered.

To prove causality, another concept must be used. In a statistical sense, such a concept is Granger causality by C. W. J. Granger (Granger, 1969). He posits a definition that if a variable can be predicted better with the inclusion of another variable than without it, then the latter causally influences the former. This is explicitly called Granger causality or predictive causality, since it cannot be considered a test for true causality[1]. If X and Y are actually caused by Z, the test of causal relationship between X and Y might still indicate causality using Granger's concept. This may be related to the problem of knowing exactly which variable is endogenous and which is exogenous (Granger, 1980).

Sims adapted this causality concept by stating that a variable X is strictly exogenous relative to another variable Y if the linear predictor of Y using all past, present, and future values of X is identical to the linear predictor based on all past and present values of X without considering the future (Sims, 1972). Thus he relaxes Granger's axiom that "the past and the present can cause the future, but the future cannot cause the past" (Granger, 1980, p. 330).

In his famous essay "Money, Income, and Causality", Sims not only established what is called Sims causality, i.e., considering the future to infer exogeneity, but also showed that Granger causality is equivalent to the parameter restrictions of the moving average representation (Sims, 1972). Later, Florens and Mouchart showed that Granger causality is a stronger condition. While Granger causality implies Sims causality, the converse is not true (Florens & Mouchart, 1982).

To return to Sims: He didn't just expand Granger's concept, but also proposed the vector autoregressive VAR model that is now the usual basis for conducting Granger causality tests. He proposed the VAR model as a tool for evaluating alternative macroeconomic models (Sims, 1980). VAR models can be used, for example, to predict economic time series, to develop and evaluate economic models, and to assess the consequences of alternative policies.

[1] True causality goes beyond the realm of statistics and is vividly discussed in philosophies (Granger, 1980).

A restriction of the VAR model is, that it relies on stationary time series, i.e., the first two moments must be time-independent. When time series are non-stationary, one faces a problem when using the VAR model. Either the data are preprocessed to make them stationary, or the method proposed by Toda and Yamamoto is used (Toda & Yamamoto, 1995). It allows to estimate the VAR model with integrated or cointegrated processes of arbitrary order by extending the lag length using an additional number d_{max} that is the maximal order of integration suspected to occur in the process. If a lag length k is determined, this leads to estimating a $(d_{max} + k)$th-order VAR model.

As for the practical application of Granger and Sims causality, Sims has shown, for example, using postwar data from the U.S., that money causes income[2], but that this relationship does not hold in reverse. Hence, indicating that money is exogenous in the sense of this relationship when applying a regression of GNP on current and past money (Sims, 1972).

Based on the publication dates of the papers discussed, it is clear that the concept of Granger and Sims causality is quite old. Recent developments extend the concept of statistical causality to take into account, for example, nonlinearity or model-free measurements. One such measurement is Thomas Schreiber's concept of transfer entropy (Schreiber, 2000), which is an information-theoretic approach that is model-free and allows for nonlinear time series. Like Granger causality, it attempts to explain the relationship between variables by quantifying the information exchanged between them. Knowing that there are more sophisticated methods for testing causality, this paper is limited to the concept of Granger causality and mainly its linear implementation to test a set of linearly generated time series.

Structure of this paper The main objective of this paper is to test different software libraries written in R and Python and compare their performance with regard to their Granger causality implementation, i.e., how well they find causal relationships and whether they are biased when time series have different properties (e.g., show a time trend, are correlated, or include structural breaks). To generate time series, the logic of the VAR model is used as a framework for the data generating process. For this reason, and due to the fact that Granger causality in its usual form is based on the VAR model, it is introduced in Section 2.1, followed by the formal concept of Granger causality in Section 2.2. The empirical part is explained in Chapter 3, where the study design as well as the different analyses performed are presented.

General information: All figures and tables in this paper are self-generated and do not come from sources other than the author's own. Therefore, the table and figure labels do not indicate a source to avoid redundant information.

[2] Money is measured using M1 data (Federal Reserve Bank of St. Louis) and income using U.S. gross national product (GNP).

2 Methodology

The following chapter is a brief introduction to simultaneous equation models in a time series context and lays the foundation for the time series simulation in Chapter 3. This simulation relys on various different parameters that are going to be explained in Section 2.1.

The subsequent Section 2.2 will then elaborate on the concept and methodology of Granger causality.

2.1 Vector Autoregressive Model

Dynamic econometric models extend the usual static regression model by either incorporating only the lagged values of the exogenous variables (distributed-lag model) or only the lagged values of the endogenous variable (autoregressive model)[1] [2].

Combining the distributed lag model and the autoregressive model leads to the autoregressive distributed lag *ADL(p, q)* model, a regression model that regresses the endogenous variable on its p lagged values and the q lagged values of the exogenous variable. If y_t can be explained by its p lags and q lags of x_t, then the model is

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots + \delta_q x_{t-q} + u_t$$

[1] Random variables are denoted by upper case letters and their realizations by lower case letters, i.e., X is a random variable with its realizations x. Both can be indexed to time: X_t and x_t, with $t \in \mathbb{T}$, where $\mathbb{T} = \mathbb{Z}$.

[2] Given two time series $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are the vectors of realized values of the stochastic processes $\mathbf{X}$ and $\mathbf{Y}$, both types can be represented by

$y_t = \beta_0 + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots + \delta_q x_{t-q} + u_t$ (distributed-lag *DL(q)* model)
$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + u_t$ (autoregressive *AR(p)* model)

For more insights, see chapter 15 in Stock and Watson, 2020, p. 568 and p. 571.

with the unknown coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_p, \delta_1, \delta_2, \ldots, \delta_q$ and the error term u_t with $\mathbb{E}(u_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}, x_{t-1}, x_{t-2}, \ldots, x_{t-q}) = 0$ (Stock & Watson, 2020, chapter 15, p. 572).

Leaving the bivariate context, one can specify a more general *ADL* model with E exogenous variables and their respective $Q_1, \ldots, Q_E$ lags as

$$y_t = \beta_0 + \sum_i \beta_p y_{t-i} + \sum_e \sum_{q \in q_e} \delta_{e,q} x_{e,t-q} + u_t,$$

where $i = 1, \ldots, p$, $e = 1, \ldots, E$ and $q_e = \{1, \ldots, Q_e\}$.

This single-equation model can then be extended to a multi-equation model, the vector autoregressive *VAR* model (Sims, 1980), where each variable is both exogenous and endogenous. Therefore, it is no longer necessary to distinguish between lags of exogenous (q) and endogenous (p) variables. Hence, lags are now denoted by p.

Given K different variables, K different ADL models arise. With $k = 1, \ldots, K$ and $i_k = 1, \ldots, p_k$, where α_{k,i_k}, ν_k and γ_k represent the coefficients, the constant and the deterministic trend of the kth time series, respectively. All error terms $u_{1t}, \ldots, u_{Kt}$ are independently and identically distributed with zero mean. The equation system reads

$$\begin{bmatrix} y_{i,t} \\ \vdots \\ y_{K,t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_K \end{bmatrix} + \begin{bmatrix} \gamma_1 t \\ \vdots \\ \gamma_K t \end{bmatrix} + \sum_k \sum_{i_k} \begin{bmatrix} \alpha_{k,i_k} y_{k,t-i_k} \\ \vdots \\ \alpha_{k,i_k} y_{k,t-i_k} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{K,t} \end{bmatrix}.$$

The next and final step is to write this multi-equation system in matrix notation. With $i = 1, \ldots, p$, $\mathbf{y}_t := (y_{1t}, \ldots, y_{Kt})^T$, $\boldsymbol{\nu} := (\nu_1, \ldots, \nu_K)^T$, $\mathbf{u}_t := (u_{1t}, \ldots, u_{Kt})^T$, $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_K)^T$ and $\mathbf{A}_i$ representing the coefficient matrix of the ith lag

$$\mathbf{A}_i = \begin{bmatrix} \alpha_{11,i} & \cdots & \alpha_{1K,i} \\ \vdots & \ddots & \vdots \\ \alpha_{K1,i} & \cdots & \alpha_{KK,i} \end{bmatrix}$$

the compact form in matrix notation (Lütkepohl, 2005, chapter 2, p. 12) is given by

$$\mathbf{y}_t = \boldsymbol{\nu} + \boldsymbol{\gamma} t + \sum_i \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t, \qquad \mathbf{u}_t \overset{d}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{2.1.1}$$

Equation (2.1.1) is highly versatile for generating time series data. Not only can one specify different coefficients for each lag matrix $\mathbf{A}_i$ to precisely control the causal influence between

the different time series, but one can also make them stationary by simply removing the deterministic trend γt, thus $\gamma = \mathbf{0}$[3].

2.2 Granger-Causality

The concept of Granger causality attempts to indicate whether the prediction error of one variable decreases by including other variables in the regression model (Granger, 1969). Considering a bivariate setting with two a stationary stochastic processes X_t and Y_t, where $\overline{X}_t$ and $\overline{Y}_t$ denote the set of past values $\{X_{t-j}, j = 1, 2, \ldots, \infty\}$ and $\{Y_{t-j}, j = 1, 2, \ldots, \infty\}$.

If $\mathcal{P}_t(X|\overline{X})$ is the best and unbiased least-squares predictor of X_t using only the past values $\overline{X}_t$, then $\sigma^2(X|\overline{X})$ is the variance of the prediction error. Furthermore, U_t contains all the accumulated information in the universe until time $t-1$ and $U_t - Y_t$ all this information apart from Y_t.

Y is then said to cause X if $\sigma^2(X|\overline{U}) < \sigma^2(X|\overline{U-Y})$. In words: If X_t is being better explained by including all information of Y_t than without it, Y_t is Granger causing X_t denoted by $Y \Rightarrow X$.

Two closely related causality concepts are *feedback* and *instantaneous causality*:

- Feedback ($Y \Leftrightarrow X$) refers to the phenomena that not only Y_t is Granger causing X_t, but X_t is Granger causing Y_t as well. In this scenario $\sigma^2(X|\overline{U}) < \sigma^2(X|\overline{U-Y})$ and $\sigma^2(Y|\overline{U}) < \sigma^2(Y|\overline{U-X})$ hold.
- Instantaneous causality describes that incorporating the current value of a variable[4] helps to improve explaining another variable. If $\sigma^2(X|\overline{U}, \overline{\overline{Y}}) < \sigma^2(X|\overline{U})$ then Y_t instantaneous Granger causes X_t.

To test for Granger causality, which is actually a test for non-Granger causality[5], the Wald test[6] can be employed. It tests the hypothesis that all coefficients on $\overline{Y}_t$ are jointly zero in the equation for X_t. It can easily be converted from a bivariate to a multivariate test, checking whether the E stationary stochastic processes in $Y_t = (Y_{1t}, \ldots, Y_{Et})^T$ Granger cause X_t. If $\boldsymbol{\alpha} = vec[\mathbf{A}_i, \ldots, \mathbf{A}_p]$ is the vector of all VAR coefficients then the noncausality restrictions can be stated as $\mathbf{R}\boldsymbol{\alpha} = \mathbf{c}$. Hence, the hyothetheses are

$$\mathbb{H}_0 : \mathbf{R}\boldsymbol{\alpha} = \mathbf{0} \quad \text{against} \quad \mathbb{H}_1 : \mathbf{R}\boldsymbol{\alpha} \neq \mathbf{0} \tag{2.2.1}$$

where $\mathbf{R}$ is the restriction matrix (Lütkepohl & Reimers, 1992).

[3] The stationary VAR model does not contain a deterministic time trend by definition, so the Equation (2.1.1) is a slightly adjusted version. With $\gamma = \mathbf{0}$ this reduces to the original VAR model equation.
[4] Denoted by a double overline.
[5] Due to the hypositheses in Equation (2.2.1).
[6] The Wald test usually reports small-sample F statistics or large-sample χ^2 statistics (see StataCorp, 2021, vargranger and Zeileis and Hothorn, 2002, waldtest).

3 Empirical analyses and performance comparisons

The methodological foundations of the VAR model and Granger causality were presented in the previous chapter in order to understand what is going on in econometric software packages that perform a Granger causality test.

The quality of those software packages is the main focus of this paper. Therefore, the goal of this chapter is to explain which packages have been selected, what are their characteristics, and how the analysis procedure is implemented in Python and R. Finally, single and sensitivity analyses are performed to graphically show the goodness of fit for all packages.

3.1 Structure of the analyses

Granger causality can be conducted in a bivariate as well as a multivariate setting. With $K = 4$ time series $\mathbf{T} = (T1, T2, T3, T4)$, a bivariate analysis can test wether $T2 \Rightarrow T1$, $T3 \Rightarrow T1$, $T4 \Rightarrow T1$, $T1 \Rightarrow T2$ etc. Hence, it tests wether $T(i \in S) \Rightarrow T(j \in S)$ for all $i \neq j$ with $S = \{1, \ldots, K\}$.

Bivariate analysis is nested in multivariate analysis, which tests not only whether one time series causes another, but also whether multiple time series cause another. For example, it tests whether $(T1, T2) \Rightarrow T4$, $(T1, T2, T3) \Rightarrow T4$, $(T1, T3) \Rightarrow T4$, and so on.

The empirical analysis therefore focuses on the two test types *bivariate* and *multivariate* with additional subtypes. A common bivariate Granger causality test depends on a lag parameter p, which specifies up to which lag the covariates are included in the regression model. For example, with $p = 4$ one can test wether $\overline{T2}_t \Rightarrow T1_t$ with $\overline{T2}_t = \{\overline{T2}_{t-j}, j = 1, 2, \ldots, p\}$. The test result then indicates whether all $p = 4$ lags of $T2$ Granger cause $T1$. To know precisely whether a particular lagged term is Granger-causal, one can perform a partial test, which thus differs in design from the *aggregated bivariate* test described above[1].

[1] See Appendix B for details.

In multivariate analysis, one can perform a complete analysis of all time series combinations or limit oneself to whether one time series is Granger-caused by all other time series, which is called *all on one* in this paper. The different test types are stated in Table 3.1.

Bivariate	**Multivariate**
Aggregated	Complete
Partial	All on one

Table 3.1: Granger causality test types

3.1.1 Packages

Packages that test on non-Granger causality[2] are limited to the programming languages `Python` and `R`. The `Python` package used is `statsmodels` (Seabold & Perktold, 2010), whereas the R packages are `vars` (Pfaff, 2008a; Pfaff, 2008b), `lmtest` (Zeileis & Hothorn, 2002), `VLTimeCausality` (Amornbunchornvej et al., 2021), `NlinTS` (Hmamouche, 2020) and `bruceR` (Bao, 2022). The partial bivariate Granger causality test is written from scratch in `Python` making use of functions provided by `statsmodels`.

With regard to Table 3.1 the packages mentioned above can be assigned to the two primary test categories with the specific function names given in Table 3.2 and Table 3.3.

Packages	Complete	Partial
statsmodels	`test_causality()`	
vars	`causality()`	
lmtest	`grangertest()`	
VLTimeCausality	`VLGrangerFunc()`	
NlinT	`nlin_causality.test()`	
bruceR		
Own implementation		`partial_granger_causality()`

Table 3.2: Python and R packages for bivariate Granger causality tests

All tests except `nlin_causality.test()` test for linear Granger causality. The nonlinear package `NlinTS`, which uses neural networks, is included to compare the results of a nonlinear

[2] Non-Granger causality refers to the H_0 of the Wald test. If it is rejected, this indicates a Granger causal influence. Thus, one tests explicitly for non-Granger causality and implicitly for Granger causality.

test with linear input data, but also to compare the results with the linear Granger tests when the data are not stationary, e.g. when the time series show a trend or structural breaks.

Packages	Complete	All on one
statsmodels	`test_causality()`	`test_causality()`
vars	`causality()`	`causality()`
lmtest		
VLTimeCausality		
NlinTS		
bruceR	`granger_causality()`	

Table 3.3: Python and R packages for multivariate Granger causality tests

3.1.2 Python implementation

All tests are performed automatically using the self-written Python package `CausalA`, which contains four main classes:

1. **`TS_Sim`** generates a predefined number of time series according to various parameters and implements Equation (2.1.1). Most parameters affect the way the lag coefficient matrices $\mathbf{A}_i$ are generated. The initial setting simulates stationary time series, but one can also generate time series that exhibit a trend and/or structural breaks and are therefore non-stationary.
2. **`Granger`** performs the various Granger causality tests of the different packages. It does this for the predefined test type and shows the results in different graphical representations.
3. **`SensA`** is the main class of `CausalA` that uses both `TS_Sim` and `Granger` to perform a sensitivity analysis. It takes an array of values for a given parameter of `TS_Sim` as input and plots the results like `Granger`, but this time aggregated.
4. **`Tools`** contains helper functions that are used across all other classes.

3.2 Empirical analysis

Focusing on the two main test types *aggregated bivariate* and *all on one*[3], the test are conducted for various parameter settings.

On the one hand single tests are conducted that means for one set of time series. Thereby conducting four main tests: with (1) default settings, (2) structural breaks, (3) a time trend γt and with (4) correlated error terms $\mathbf{u}_t$.

On the other hand, sensitivity analyses are performed. Each sensitivity analysis focuses on one parameter of `TS_Sim`. In an iteration process, Granger tests are performed on the newly generated time series data. All results are collected in a panel for subsequent aggregate graphical analysis[4].

Before conducting the single or sensitivity analysis, one has to generate a set of time series[5]. All subsequent analysis are conducted on $K = 5$ time series[6] that contain $T = 1,000$ time observations per time series[7]. The initial time series are stationary and shown in Figure 3.1.

3.2.1 Single analysis

The objective of this section is to perform Granger tests for different stationary and non-stationary time series to compare the performance of all packages. To do so, two main indicators are used.

Accuracy The accuracy shows how many causal influences and non causal influences are accurately detected by the packages. The accuracy of detecting the causal influences and non causal influences between the time series is presented in terms of confusion matrices and the accuracy rate[8]. Both are recorded separately for each lag[9] and package.

[3] The test environments of *aggregated bivariate* and *all on one* also contain tests for *partial bivariate* and *complete multivariate*, but since only a small number of packages allow testing for the latter two, they play only a minor role.

[4] Due to space limitations, not all illustrations are included, but they are in the `empirical results` folder, which is attached to this document. Also, running the `.ipynb` files will automatically generate all the figures and can be used to run further tests with different parameter settings.

[5] Please see Appendix A.4 on how the coefficients are generated.

[6] The use of five time series allows a good interpretability of the matrices of lag coefficients and at the same time allows a realistic investigation of multiple time series. In addition, as the number of time series increases, computational time and capacity rise sharply (especially in the true multivariate case). Therefore, five time series seem to be a good compromise. The analysis of additional time series can be conducted with ease using the attached jupyter notebook file.

[7] `seed=128` is used.

[8] See Appendix A.3 for more information.

[9] Except in the case of *partial bivariate*, the *aggregated bivariate* and *all on one* tests show the aggregated detected causal influences for the specific lags. For details on how the detected causal influences are aggregated, see Appendix A.

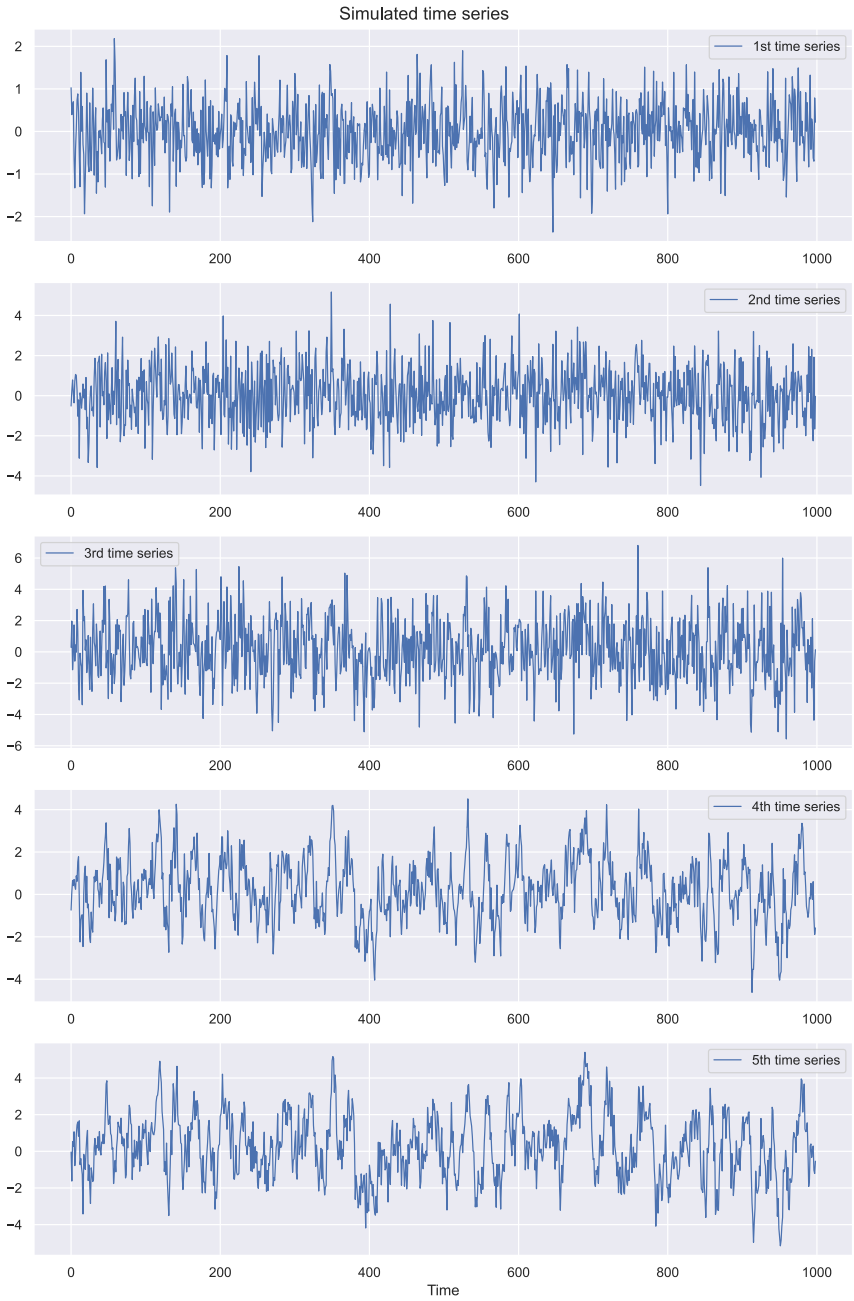


Figure 3.1: Five simulated time series according to the default parameters

In addition, a comparison between the *aggregated bivariate* and *partial bivariate* Granger causality tests should show whether they differ in terms of their accuracy rate. In the case

of true multivariate Granger tests, the accuracy rate has to be approximated[10].

Correlation Correlation tests are performed to determine whether high true coefficients lead to low p-values of the Granger test statistic and whether these p-values correlate with the true causal influences[11]. For the first question, a common Pearson correlation coefficient is calculated, while for the second question, the point biserial correlation coefficient is used because the true causal influence is binary coded. The p-values of the correlation tests also allow to distinguish between significant and non-significant correlation coefficients.

Pearson correlation coefficient

Question: Do high true coefficients lead to low p-values of the Granger test statistic and vice versa?

The true causal influences $\mathbf{C}_i$ may even contain ones for very low true coefficients. This is due to the way the time series are simulated. In Granger tests, this can lead to biased results if, for example, the coefficients loose strength with each lag and the test statistic is thus mainly driven by the high coefficients of prior lags. However, the Granger test checks that all coefficients are Granger-causal up to the desired lag. If the coefficients of the first few lags are high, this may result in low p-values, even if the coefficients of the later lags are very small.

In general, high true coefficients should lead to low p-values of the Granger tests. Hence, the Pearson correlation should be strongly negative. If now lag terms are added to the regression model that show low true lag coefficients and the test statistics of the Granger tests do not rise, this will lead to a decrease of Pearson correlation coefficient. Indicating that it might be better to exclude the newly added lag terms from the regression model because they are not that relevant in terms of their explanation power.

Point biserial correlation coefficient

Question: Do true causal influences lead to low p-values of the Granger test statistic and true non-causal influences to high p-values?

When estimating the causal influences $\hat{\mathbf{C}}_i$, we do not distinguish whether the p-values are close to zero or close to the significance level. The only important thing is that they are

[10] See again Appendix A.3 for discussion and calculation examples.
[11] See Appendix A.2 for details.

smaller than it. In this way, the information about the estimation certainty of the test is lost. Therefore, the accuracy rate and the confusion matrices are only half perfect[12].

Another way to measure the accuracy is to calculate the correlation between the p-values of the Granger tests and the true causal influences. Thus, a high negative correlation would mean that the test can clearly distinguish between causal and non-causal influences. Causal influences would ideally show p-values close to zero and non-causal influences would show p-values close to one. From this point of view, the point biserial correlation coefficient indicates the discriminatory power of the Granger test.

3.2.1.1 Stationary time series

The first single analysis block is concerned with stationary time series, i.e., time series that have a probability distribution that does not change over time. This property is important for regression models that use past data to predict future values (see Stock and Watson, 2020, chapter 15).

Time series with default parameters[13] Time series according to the initial parameter settings as shown in Figure 3.1 are used to perform Granger causality tests. Starting with the aggregate bivariate tests, we are interested in whether there is a bivariate relationship between them, that is, whether one time series can be explained by the lagged terms of another.

Aggregated bivariate tests The absolute and relative confusion matrix for the first lag of `statsmodels` is shown in Figure 3.2[14] indicating that eight true negative and ten true positive events are detected. Translating to an accuracy rate of 90 %.

The accuracy rates for all seven lags are displayed in Figure 3.3. The blue line gives an orientation about the strength of the decrease of the true coefficients A_i. The intuition behind this is that the lower the true coefficients, the more difficult it is assumed to estimate the causal influence by the Granger test.

After specifying the confusion matrices and accuracy rates, the correlation plot in Figure 3.4 makes it clear that the *aggregated bivariate* Granger test of `statsmodels` is significantly negatively correlated with the true coefficients at all lags. This is true for both correlation

[12] Example: Having two p-values of 0.001 and 0.026 and a significance level of 5 %. If this result leads to a perfect accuracy rate and the p-values rise to 0.039 and 0.046, the accuracy rate would still be perfect, but the test has clearly lost discriminatory power, which we cannot deduce from the accuracy rate and the confusion matrix.

[13] In this very first analysis, the test results and diagrams are discussed in detail to explain how these illustrations can be interpreted. However, since this type of discussion is too lengthy, limitations must be made afterwards.

[14] All confusion matrices for remaining lags are illustrated in Figure D.3.

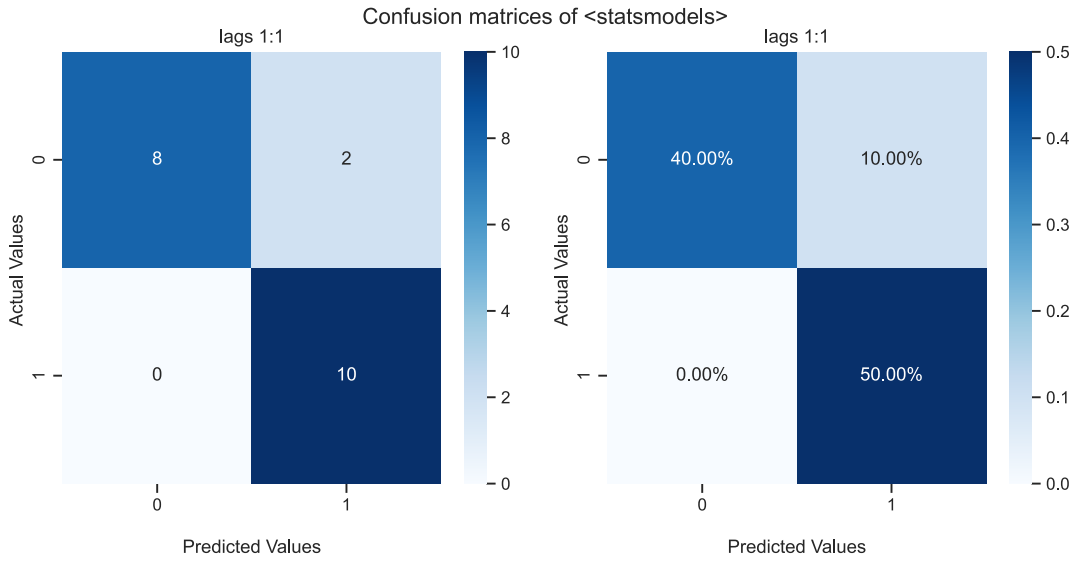


Figure 3.2: Confusion matrices for the first lag of `statsmodels` (excerpt)

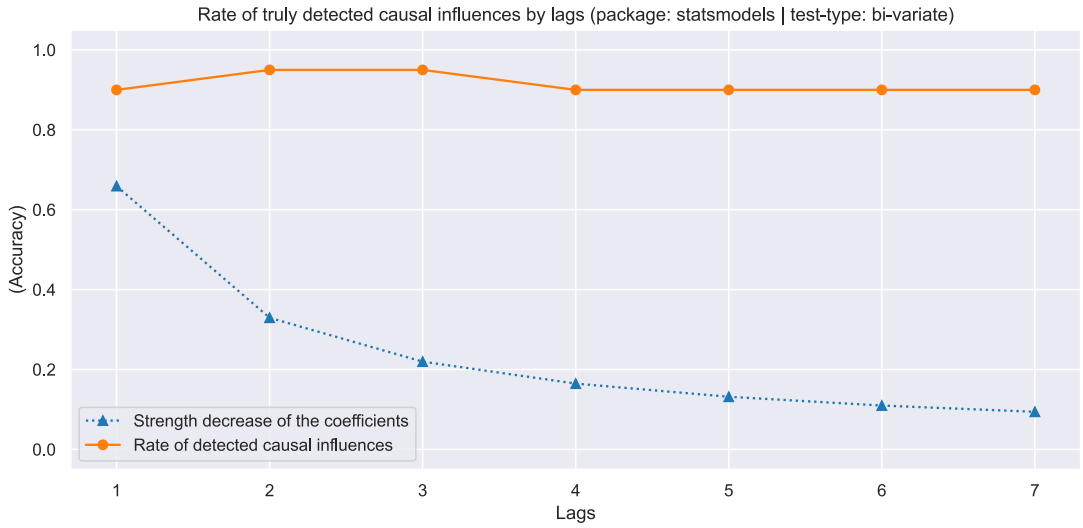


Figure 3.3: Accuracy for the first lag of `statsmodels`

coefficients. The correlation coefficients are strong for the first lags, but show a decreasing trend, indicating that the more lags are included in the Granger test, the less the p-values represent the influence of the true coefficient in terms of their magnitude and true causal influences. The test loses discriminatory power and suffers from bias due to the higher true coefficients of the very first lags.

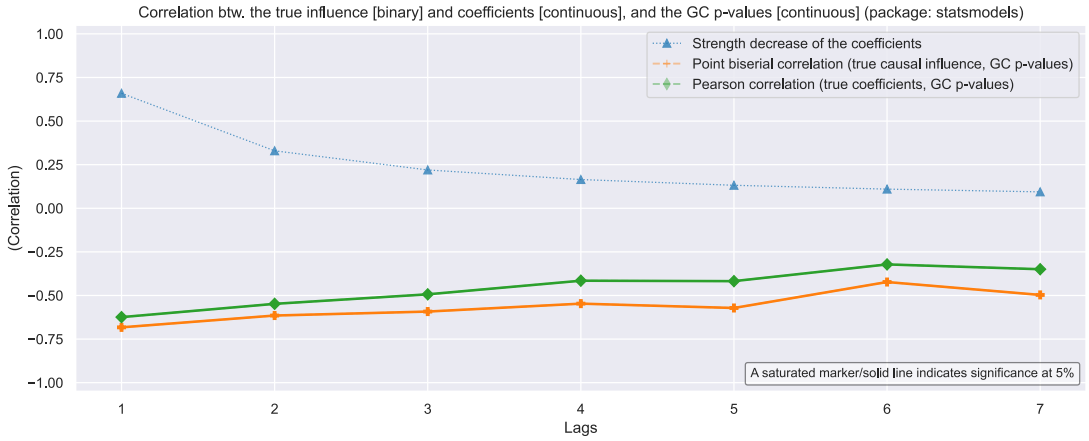


Figure 3.4: Correlation plot of `statsmodels`

As a result, `statsmodels` detects the true causal and non-causal influence quite well when referring to the accuracy rate, but shows decreasing correlation coefficients. This means that `statsmodels` is not able to clearly separate causal from non-causal coefficients and that it is biased by previous lag terms.

Partial bivariate Comparing the *aggregated bivariate* results to the partial bivariate results shown in Figure 3.5, makes it clear that aggregation of lag terms leads to an enhanced accuracy rate up to 40 percentage points. This result holds for all other packages except `NlinTS`, which does not perform sufficiently better than the partial Granger causality test. It appears that this "memory" of the *aggregated bivariate* test is superior to a partial focus on a single lag term. It is therefore preferable to include more lags in the regression model and thus aggregate them.

Complete multivariate In the *true multivariate* setting, all combinations of time series except the one tested are considered to Granger cause the latter. Accuracy is therefore approximated by comparing the mean of estimated causal influences per time series with the mean of the true causal influences for the same time series. Figure 3.6[15] therefore shows that all packages overestimate the causal impact on the second time series among all lags.

[15] Figure D.4 includes all plots.

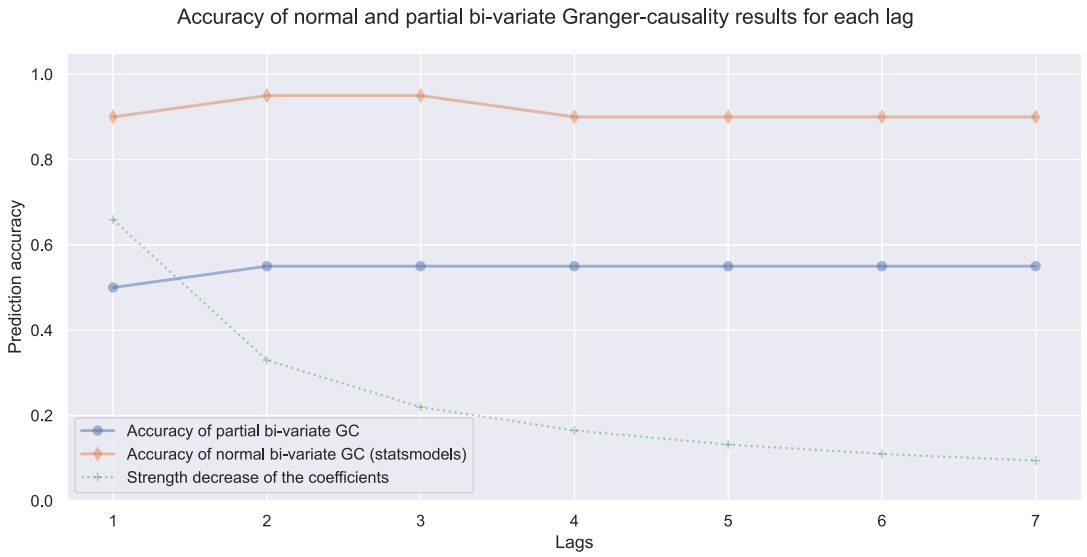


Figure 3.5: Comparison between the aggregated and partial bivariate Granger test <`statsmodels`> (excerpt)

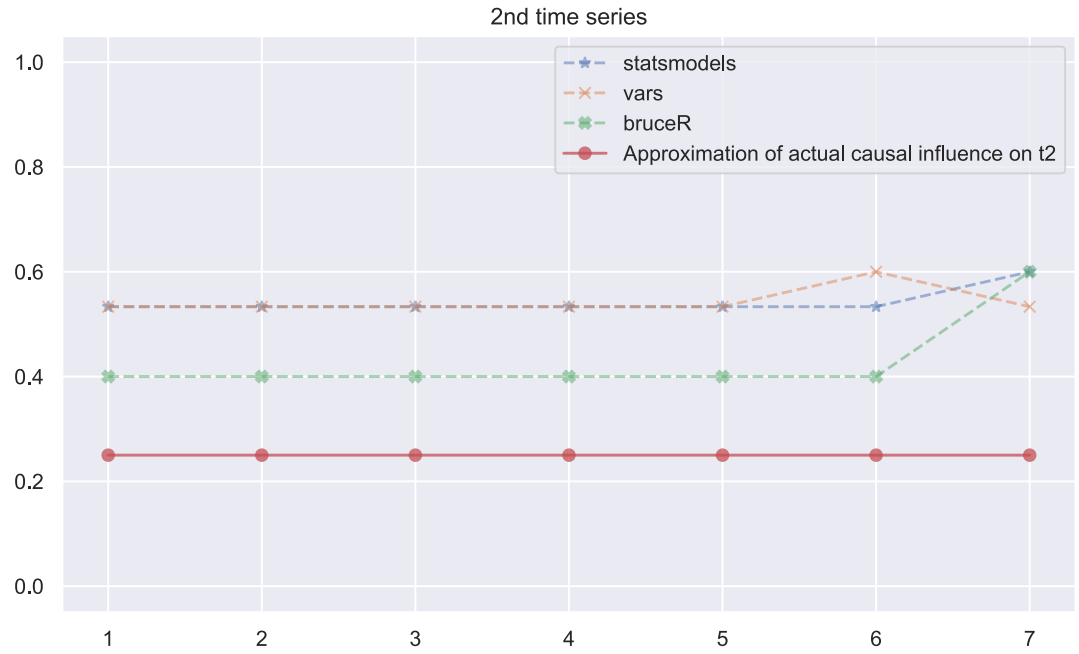


Figure 3.6: Approximated accuracy per time series in the *true multivariate* case (excerpt)

All on one In the *all on one* setting, the relevance indicator (Pearson correlation) has decreased slightly from -45 % to -30 %, reflecting the decrease in coefficient strength. Hence, the coefficients of the first lag appear to bias the test statistics. However, the discriminatory

power (point biserial correlation) remains constant at -50 %, indicating that with newly added lags, the correlation between the p-values and the true causal influences does not change. Thus, the p-values do not increase or decrease sufficiently[16].

From this and other diagrams it can be seen that the correlation coefficients hardly take values close to -1. Looking at the p-values[17] of this example, we see that all time series except the first are estimated to be caused by all others. The p-values are close to zero, but the p-values of the first time series, with one exception, are not significant, i.e., greater than the significance level of 5 %.

By definition, the first time series is not caused by the others. A Granger test that truly captures this with certainty should have a high p-value close to or even at 1. However, this is not the case in this example. The respective p-values are not significant, but they are not really high either. Consequently, the correlation cannot reach a strong negative correlation, although the test proves causality with high certainty, but not non-causality, at least not with the same strong certainty.

Summarized results From the bivariate analysis one clearly sees that the `VLTimeCausality` Granger test is superior to all others reaching an accuracy rate of 100 %. In an *all on one* comparison the accuracy of `statsmodels` and `vars` is constant at 70 % among all lags. Looking at the confusion matrices, it is clear that both show no false negatives, which means that the test is perfect when it comes to detecting true causal influences, but shows weaknesses when it comes to detecting true non-causal influences.

`NlinTS` and the `partial Granger` test perform the worst. The confusion matrices of `NlinTS` reveal, that it does not predict positives at all. Having 50 % of true causal influences by definition[18] thus gives a constant accuracy rate of 50 % when only predicting negatives.

The correlation analysis shows that `VLTimeCausality` has the highest negative correlation for both types. It decreases for each additional lag, but always remains significant. `lmtest/vars`[19] and `statsmodels` also show decent results, but fall short of `VLTimeCausality`. All four show decreasing correlation coefficients, indicating that they lose discriminatory power and suffer from bias due to the higher true coefficients of the very first lags.

[16] See the corresponding diagrams in the attached folder.
[17] See the respective Jupyter Notebook.
[18] This is due to the setting `tril=True` and `coeffs_type` not being `Random`.
[19] The documentation of `vars` does not explicitly state that it uses the `lmtest` Granger test. All subsequent tests show mostly the same results for both libraries. Looking at the console output when loading the `vars` package in `R`, one reads the following line `<loading required package: lmtest>`. Therefore, it is assumed that `vars` uses the `lmtest` Granger test. Therefore, both are considered as one package.

Necessary restrictions In the following analysis, the focus is on the Pearson correlation coefficient for the following two reasons:

1. The accuracy rate is calculated from the confusion matrix comparing the true with the estimated causal influences. The former has the disadvantage that even very small true coefficients are indicated by a 1 in the true causal influence matrices, and the latter has the disadvantage that if the p-value is below the significance level it is decoded by 1 otherwise by 0. Thus, one loses the information contained in the continuous nature of the p-values. Furthermore, the accuracy rate can be misleading if a test predicts only true values while the actual values are half true and half negative. The accuracy rate is then 50 %, but the test is far from good. Not only is the accuracy of 50 % not optimal, but the reason is even more dramatic. The test is not able to display negatives, which should be a major concern.

2. The point biserial correlation gives an indication of the discriminatory power, but this measure targets the true causal influence matrices, which, as mentioned earlier, are not bias-free because even small coefficient are indicated by a 1 in the matrices $\mathbf{C}_i^{init}$.

Therefore, from now on, only the Pearson correlation coefficient will be used to provide a more multi-faceted analysis by comparing two continuous numbers, the true coefficients with p-values. To further simplify the comparison, Pearson correlation coefficients are summarized over the range of all lags per package, and presented in box plots to simplify the process of deciding, which package(s) perform best[20] [21] [22].

Another limitation is that we focus mainly on the *aggregated bivariate* results. Therefore, the partial bivariate and true multivariate results are omitted. Unless otherwise stated, box plots for the results of the *all on one* test are not presented because `statsmodels` implements the Granger test of `lmtest` and is therefore not as informative as the *aggregated bivariate* test, which contains more packages. Nevertheless, the results are discussed.

For the results with the default time series in the bivariate case, the box plots[23] are shown in Figure 3.7[24]. Supporting the analysis from above that `VLTimeCausality` performs best, thus shows a high to medium negative correlation between the p-values and the true coefficients. `lmtest/vars` performing better than the overall mean, but show a less strong correlation compared to `VLTimeCausality`. This is also true for `statsmodels`, which varies slightly more than `lmtest/vars`.

[20] The p-values of the correlation coefficients, which indicate significance, are also not used in the box plots.
[21] The full correlation plots are still available in the attached `figures` folder.
[22] With $p = 7$, seven correlation coefficients are calculated per package. These are then summarized in the box plots. The price paid for this is a loss of information that is offset by better comparability.
[23] With the blue dashed line indicating the overall mean.
[24] Table D.1 shows the corresponding descriptive statistics.

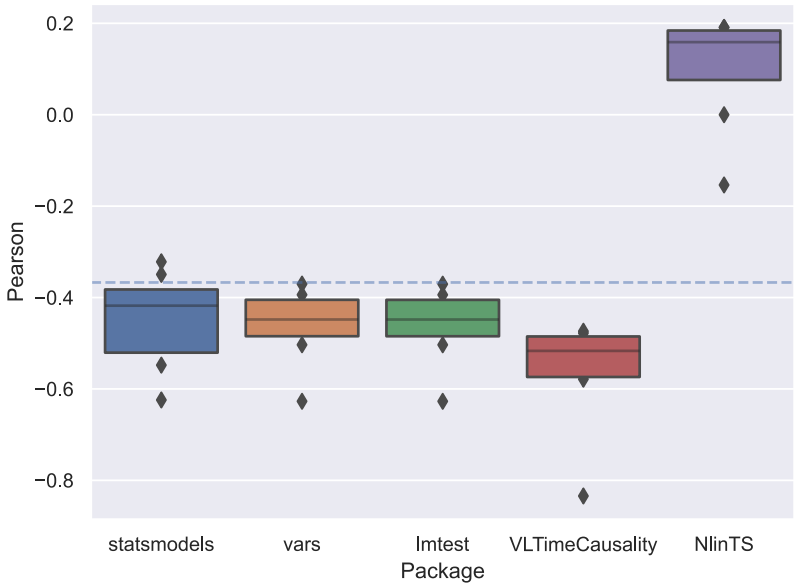


Figure 3.7: Box plots with all Pearson correlation coefficients summarized (bivariate | time series with default settings)

Correlated time series If the error terms in $\mathbf{u}_t$ of Equation (2.1.1) are correlated[25], the performance of all packages decreases, as Figure 3.8[26] shows. `VLTimeCausality` performs best and is strongly superior to all other packages. But `lmtest/vars` show the smallest deviation. The correlation is more or less stable across all lags. Therefore, both contribute to decreasing true coefficients by showing higher p-values without being biased by previous lags. Nevertheless, `VLTimeCausality` is preferable, although the correlation coefficients are not as stable as for `lmtest/vars`, but stronger, especially for the very first lags.

The *all on one* tests reveal that `statsmodels` and `vars` have quite stable correlation coefficients between all lags, like `lmtest/vars` in the *aggregated bivariate* tests. Nevertheless, `VLTimeCausality` is also preferred over `statsmodels` and `vars` because it has stronger negative correlation coefficients, especially for the first lags.

3.2.1.2 Non-stationary time series

Non-stationary time series are characterized by a mean and a variance, both of which change over time. Moreover, the value of the covariance between two time periods depends on the actual time at which the covariance is calculated, not just on the distance or gap between the two time periods (see Gujarati and Porter, 2009, Chapter 21, p. 740).

[25] $u_t \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix when the error terms are uncorrelated. In the correlated case, $\boldsymbol{\Sigma}$ contains the covariances in the cells outside the main diagonal.

[26] Table D.2 shows the corresponding descriptive statistics.

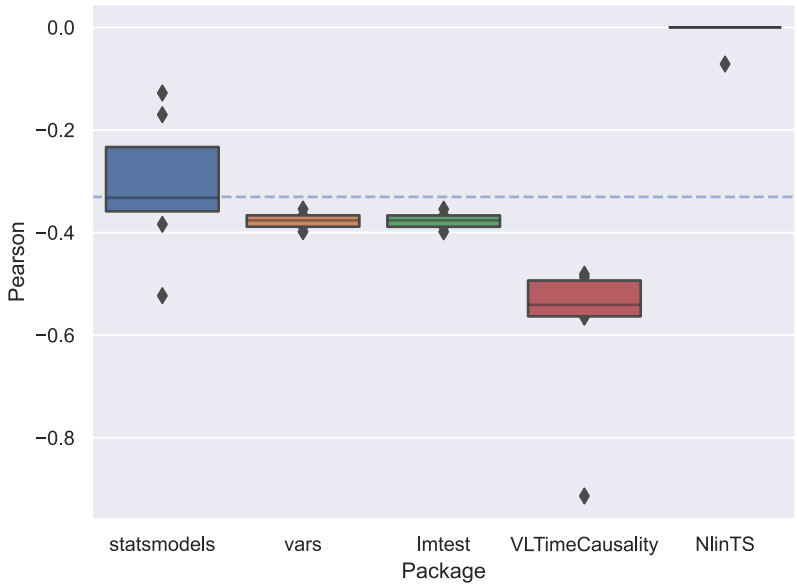


Figure 3.8: Box plots with all Pearson correlation coefficients summarized (bivariate | correlated time series)

In the following, non-stationary time series are generated featuring structural breaks[27] and incorporating a deterministic trend. Time series with structural breaks are illustrated in Figure D.1 and time series with a time trend are visualized in Figure D.2.

Time series with structural breaks The corresponding time series are simulated to include three structural breaks at random time points. Not all five time series contain structural breaks, but only two are randomly selected[28].

Figure 3.9[29] indicates that the performance does vanish in comparison to stationary time series (see Figure 3.7). `VLTimeCausality` shows the best performance thus being more sensitive to the true coefficients values. `NlinTS` has not performed well in the stationary cases, but even shows no advantage when using time series with structural breaks. `vars` and `lmtest` perform worse than the mean and even differ from each other, so the conclusion that `vars` implements the Granger causality test of `lmtest` can be called into question[30].

In the *all on one* setting, `statsmodels` and `vars` perform differently, showing a mean correlation coefficient of -0.31 and -0.21 with a standard deviation of 0.036 and 0.14, respectively.

[27] Thus, the mean vector and the covariance matrix of the error term may suddenly change to a greater or lesser extent.
[28] The number of time series with structural breaks varies with other seeds because it depends on the random number generator.
[29] Table D.3 shows the corresponding descriptive statistics.
[30] At least for times series that show structural breaks.

Thus, this type of Granger test is again not superior to the *aggregated bivariate* test.

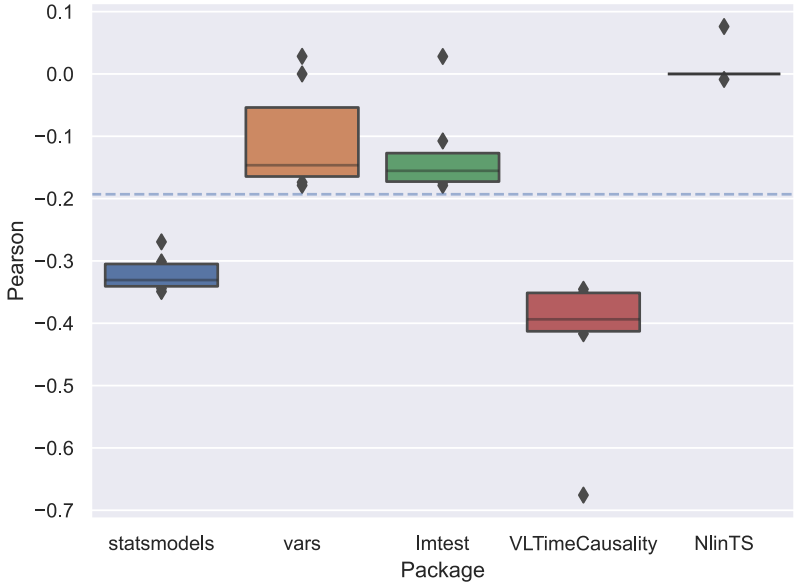


Figure 3.9: Box plots with all Pearson correlation coefficients summarized (bivariate | time series with structural breaks)

Time series that show a time trend Figure 3.10[31] makes it clear that all packages have massive problems indicating causality when the time series are not stationary in the sense of a time trend. This is not surprising, since Granger tests relying on the VAR model assume stationary time series by definition (see Toda and Yamamoto, 1995).

The *all on one* test shows that the `vars` package breaks, while `statsmodels` has a mean correlation of -0.31 with a small standard deviation of 0.036[32]. Thus, the performance of *all on one* is superior to the bivariate performance for non-stationary time series showing a deterministic trend.

3.2.2 Sensitivity analysis

Having performed Granger tests on a selection of different time series types, it is now of great interest to see how performance changes as we iterate over a given parameter of `TS_Sim`.

[31] Table D.4 shows the corresponding descriptive statistics

[32] The values of the summary statistics are taken from the `Pearson_corr_summary.txt` in the corresponding folder.

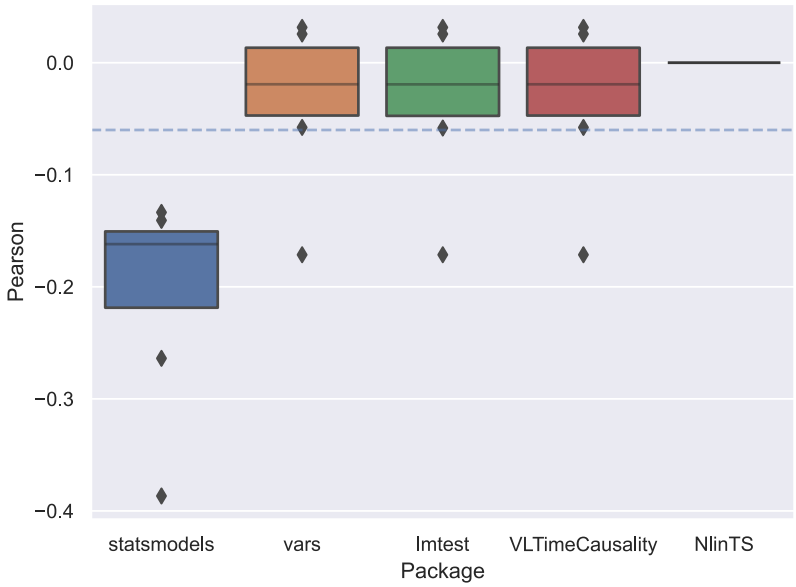


Figure 3.10: Box plots with Pearson correlation (bivariate | time series with a time trend)

Thence, this section focuses on sensitivity analyses[33] targeting the true coefficient matrices (Section 3.2.2.1) and the error terms (Section 3.2.2.2). The detailed diagrams of the sensitivity analyses are explained in Appendix C[34], as we will continue to focus mainly on the Pearson correlation, using boxen plots[35] that allow for a larger number of quantiles and thus can provide visually more detailed insights into the distribution of the data.

3.2.2.1 Variation of the coefficients

Decay strength The first sensitivity analysis targets the decay strength s of Equation (A.0.1). A single Granger causality analysis over all seven lags is conducted for each value in the set $s = \{1.00, 1.22, 1.44, 1.67, 1.89, 2.11, 2.33, 2.56, 2.78, 3.00\}$.

For the *aggregated bivariate* tests the results are summarized in Figure 3.11. This illustration confirms the previous results of the single analyses that `VLTimeCausality` performs the best, whereas apart from `NlinTS` all other packages perform more or less equally well. The quantiles below the median are longer than those above the median. Thus, most of the outliers are strongly negative coefficients, which are desirable, and the rest are more centered around the median.

[33] Considering only the *aggregated bivariate* and *all on one* type.
[34] The results of each sensitivity analysis are stored in an individual panel regarding the accuracy rates and correlation coefficients, and are located in the attached folder.
[35] The box plots are derived in the same way as the box plots of the single analysis. One takes all correlation coefficients per package to visualize them in a summarized form by the respective boxen plot.

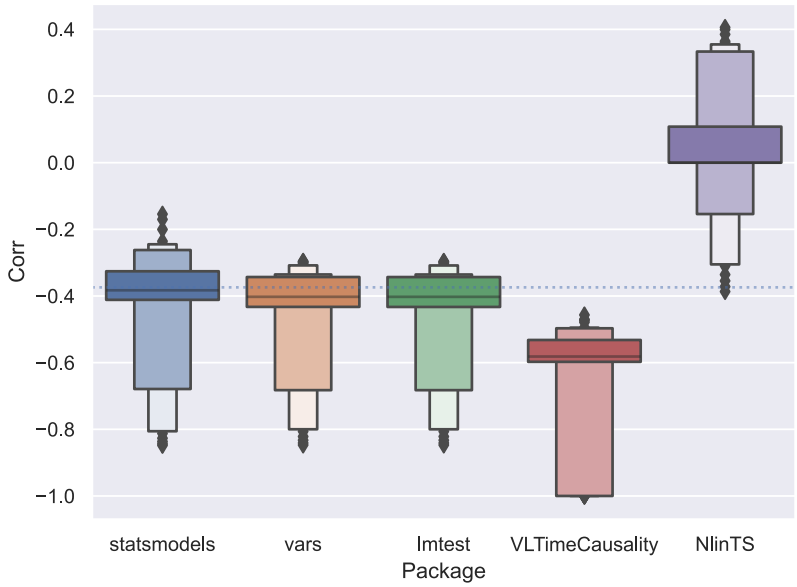


Figure 3.11: Boxenplots with all Pearson correlation coefficients summarized (bivariate | sensitivity analysis | coefficient decay scaling)

In the *all on one* setting, both packages again show the same performance, a constant correlation coefficient of -50 % across all decay strength values. Hereafter, not being superior to `VLTimeCausality`.

Evenly upward-scaling of time series specific coefficients In order to scale the coefficients of each time series upward and thus make them more uniform, the α value of the Beta distribution is increased (see Appendix A.4). The results of the sensitivity analysis are shown in Figure 3.12, with α being set to any other value between 2 and 20.

Again, `VLTimeCausality` shows the best correlation performance, especially for the first lag. `lmtest` and `vars` are clearly better than `statsmodels` this time and are mostly below the overall mean.

Looking at the performance of `statsmodels` and `vars` in the *all on one* setting, the correlation coefficients are constant at -50 % across all lags. This tells us that the relationship between the p-values and the true coefficients remains unchanged. Thus, it is insensitive to the structure of the values inside the true coefficient matrices, i.e., whether those approach the value of the autoregressive coefficients[36] or are more widely distributed.

[36] see Appendix A.4.

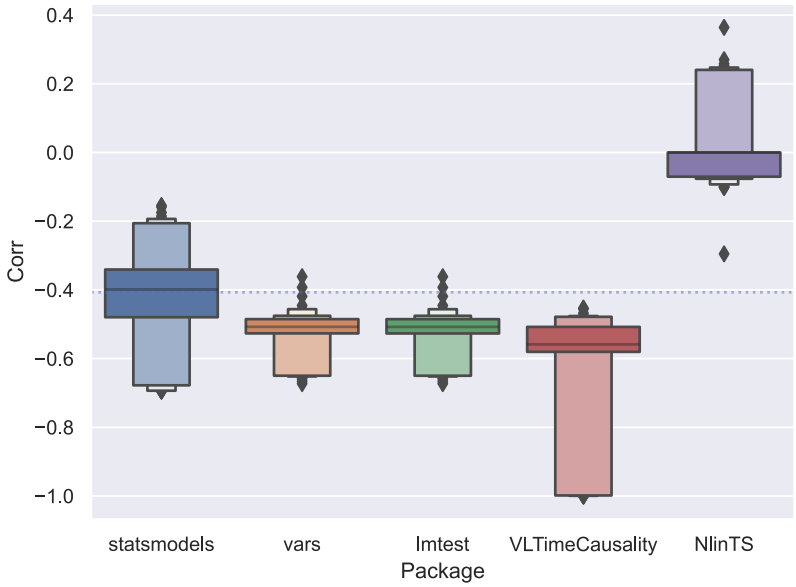


Figure 3.12: Boxenplots with all Pearson correlation coefficients summarized (bivariate | sensitivity analysis | alpha scaling)

Evenly downward-scaling of time series specific coefficients By scaling up the β of the beta distribution, the coefficients become more uniform, but now approach zero. The correlation results decrease slightly as seen in Figure 3.13 compared to Figure 3.12. But the ranking stays the same. Hence, `VLTimeCausalcity` shows the best results and `NlinTS` the worst.

The *all on one* test was insensitive to adjustments of the initial α values. The same is true for increasing β values. Thus, we see that the value structure of the true coefficient matrix has no influence on the p-values, probably because of the larger information used compared to the *aggregated bivariate* tests. When more lag terms are used, the influence of a single lag term diminishes.

3.2.2.2 Variation of the error terms

Another objective of the sensitivity analyses are the error terms. That is, increasing the number of breakpoints and scaling up the covariances in $\boldsymbol{\Sigma}$.

Increasing number of structural breaks The iteration process starts with $n_{sb} = 1$ structural breaks and goes up to $n_{sb} = 10$ with $n_{sb} \in \mathbb{N}_0$. The results shown in Figure 3.14 support the main results in the single analysis (see Figure 3.9). `VLTimeCausality` performs

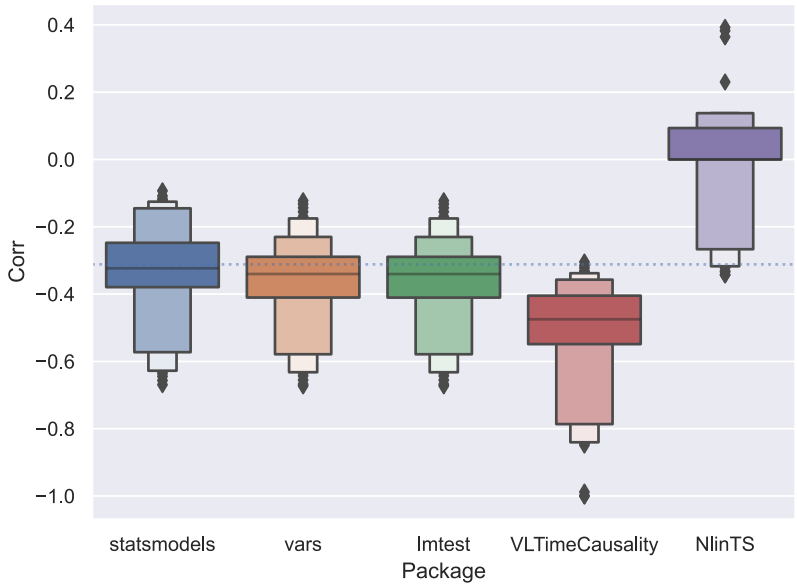


Figure 3.13: Boxenplots with all Pearson correlation coefficients summarized (bivariate | sensitivity analysis | beta scaling)

best with a mean correlation over all lags of -0.58 and a standard deviation of 0.15[37], while `NlinTS` is again underperforming.

Compared to `VLTimeCausality`, the quantiles above the median of `statsmodels`, `vars` and `lmtest` are longer than those below the median. Thus, the packages have an additional weakness, namely more correlation coefficients approaching zero.

The *all on one* analysis shows that `statsmodels` and `vars` perform largely the same across all lags, with a constant correlation of -50 %, as already evident from previous analyses. This constant correlation results from the fact that the p-values, apart from the test statistics of the first time series, are very low, close to zero or even zero. They vary, but this variation is so small that it does not alter the correlation coefficient.

A detailed view of the p-values reveals that the correlation coefficient is not close to -1 only because the tests for the first time series sometimes show low and significant p-values and sometimes show higher, nonsignificant p-values. However, the true coefficients are all zero for the first time series. An optimal test statistic would indicate this by a p-value of (close to) one. Since this is not the case, the overall correlation coefficient is weak. Consequently, the *all on one* tests have difficulty showing a non-causal relationship.

[37] The values of the summary statistics are taken from the `Pearson_sens_corr_summary.txt` in the corresponding folder.

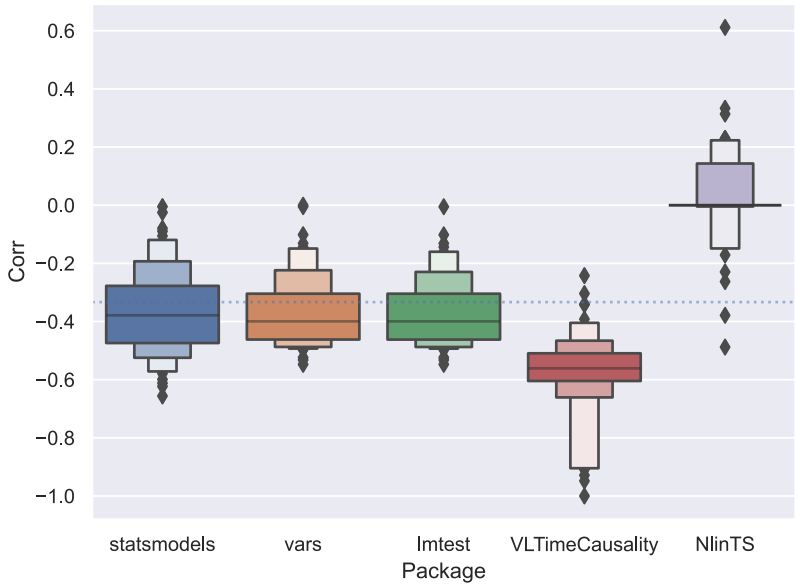


Figure 3.14: Boxenplots with all Pearson correlation coefficients summarized (bivariate | sensitivity analysis | increasing number of structural breaks)

Upscaling of the covariances In the last analysis, the covariances are scaled up. From zero, leading to no correlation at all between the time series, and reaching a scaling factor of 100. The step size is ten. As Figure 3.15 illustrates, `VLTimeCausality` shows the strongest negative correlation with a mean of -0.16 and a standard deviation of 0.20.

Interestingly, all packages except `NlinTS` show a significant decrease between a covariance scaling of 0 and 10 in terms of accuracy rate (see Figure D.6) and correlation (see Figure D.5). Disregarding `NlinTS` because its past performance was strange rather than reliable, we find that the correlation between the time series leads to biased Granger causality test statistics.

Looking at the *all on one* results, this curiosity is also visible in the correlation diagram (see Figure D.7). However, from a covariance scaling of 10, we find a downward trend with a sharp drop between 50 and 60, resulting in a correlation coefficient of about -0.6 with a covariance scaling of 100.

Thence, highly correlated time series improve the test decision of the Granger test, while only slightly correlated ones dilute it. The accuracy plot even shows an increased accuracy rate when using correlated time series (see Figure D.8).

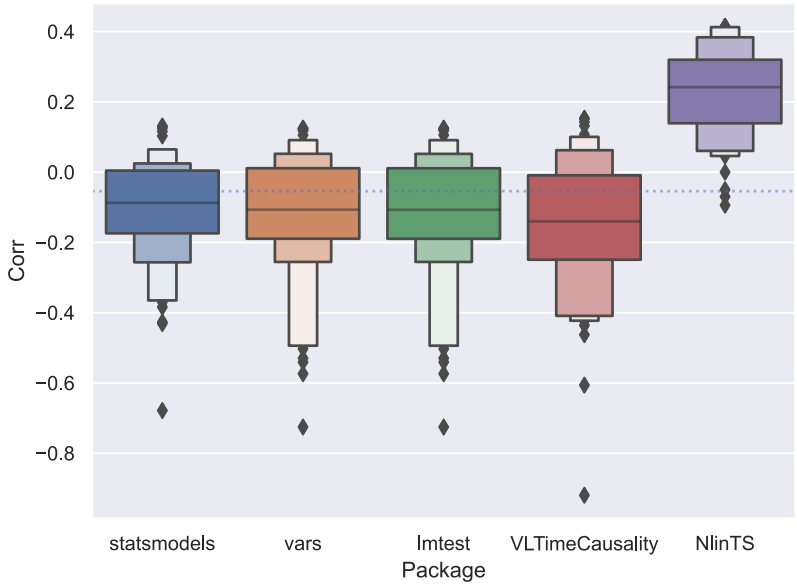


Figure 3.15: Boxenplots with all Pearson correlation coefficients summarized (bivariate | sensitivity analysis | covariance upscaling)

4 Conclusions, discussion and prospects

As a short summary, `VLTimeCausality` performs the best in most of the bivariate cases, whereas `NlinTS` performs the worst. The latter could be due to the fact that a basic setting with a single-layer neural network is used. To get a true comparison, one must first tune the hyperparameters of `NlinTS` and then compare the best model with all other packages.

The bivariate Granger tests are well suited in a stationary environment but have difficulties with non-stationary time series. Especially, if time series that have a deterministic trend, the tests break apart. This might be a good research aspect to implement the Toda and Yamamoto test (Toda & Yamamoto, 1995), which allows for integrated and thus non-stationary time series by definition.

As for correlation, the statistics of the Granger test in the *aggregated bivariate* and *all on one* setting can be misleading. The analyses performed do not provide a clear enough picture to say that one is obviously superior to the other. To be able to say this, further analyses should be performed to get a more accurate picture.

In addition, one must consider the problem of information overload when testing *all on one*. Therefore, in general, I believe that the bivariate Granger causality approach is better suited to capture the underlying relationship between two variables, i.e., whether it is uni- or bidirectional. *All on one* tests do not provide such a clear signal due to information overload.

A preference ranking for the bivariate Granger test derived from the tests performed is given by[1]:

$$\texttt{VLTimeCausality} \succ \texttt{vars/lmtest} \succeq \texttt{statsmodels} \succ \texttt{NlinTS}$$

`VLTimeCausality` is strongly preferred over all other tests. Nevertheless, `vars/lmtest` and `statsmodels` also perform well, with the former being weakly preferred due to their usually

[1] The preference ranking summarizes all the information available in the box(es) plots. Here, the ranking is not derived numerically but analytically by visually summarizing the results of the box(es) plots.

slightly better performance, except with non-stationary time series. But this is not a serious problem, as in this case another causality approach like the mentioned Toda and Yamamoto test should be used as the first option.

The analyses can be further extended by increasing the sample size T, the number of time series K, and the number of lags p among the numerous other variable parameters of `TS_Sim` to see if the results and this ranking still hold.

Another approach for comparing these packages is to select the estimated coefficients of the VAR model according to the results of the Granger test. In this way, a predictive model is created. The derived predictions can then be evaluated against the true values of the time series to measure the prediction performance. A simple metric such as the mean squared prediction error or information criteria like AIC or BIC could be used.

Additional remarks:

- *Additional individual and sensitivity analyses were performed for all Granger test types. The results are in the attached `empirical results` folder. Due to the limitations of the scope of this paper, they could not be discussed here.*
- *Furthermore, all results were performed with randomly generated values, but corresponding to a specific seed. Therefore, the results may be different for different seeds. Omitting the seed and then performing an even larger set of analyses should lead to even more meaningful results.*
- *With regard to the partial bivariate one should better use a more sophisticated concept such as the partial Granger causality test by Guo et al., 2008 that accounts for correlation.*
- *The true multivariate Granger causality test performed in this paper is unusable from the author's point of view. Due to the large number of combinations and thus test statistics, it only allows an approximate result and is therefore not very meaningful.*
- *Finally, the all on one test deserves a closer examination with respect to the influence of information overload. As the number of lag terms increases, the influence of the individual lag terms decreases. In addition, a closer look at the bias induced by strong coefficients of the first lags and this introduced "memory" is needed to see how strong this "memory" bias is compared to the aggregated bivariate case.*

A Calculation examples

Example with $K = 3$ time series, $p = 2$ lags, and initial coefficient matrices $\mathbf{A}_i^{init}$ generated by `TS_Sim` with $i = 1, \ldots, p = 1, 2$:

$$\mathbf{A}_1^{init} = \begin{bmatrix} 0.50 & 0.40 & 0.30 \\ 0.40 & 0.60 & 0.35 \\ 0.30 & 0.37 & 0.45 \end{bmatrix}$$

$$\mathbf{A}_2^{init} = \begin{bmatrix} 0.43 & 0.34 & 0.25 \\ 0.36 & 0.46 & 0.26 \\ 0.18 & 0.25 & 0.38 \end{bmatrix}$$

The causal influence is given by the matrices $\mathbf{C}_i^{init}$, which are lower triangular matrices with all elements on the main diagonal set to zero, to remove the autoregressive influences[1].

$$\mathbf{C}_1^{init} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\mathbf{C}_2^{init} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

Applying the Hadamard product $\mathbf{A}_i = \mathbf{A}_i^{init} \odot \mathbf{C}_i^{init}$, we obtain the coefficient matrices $\mathbf{A}_i$ used in the process of data generation:

$$\mathbf{A}_1 = \begin{bmatrix} 0.50 & 0.40 & 0.30 \\ 0.40 & 0.60 & 0.35 \\ 0.30 & 0.37 & 0.45 \end{bmatrix} \odot \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0.40 & 0 & 0 \\ 0.30 & 0.37 & 0 \end{bmatrix}$$

[1] This is the default setting of `TS_Sim` with `coeffs_type="Only DL"` and `tril=True`.

$$\mathbf{A}_2 = \begin{bmatrix} 0.43 & 0.34 & 0.25 \\ 0.36 & 0.46 & 0.26 \\ 0.18 & 0.25 & 0.38 \end{bmatrix} \odot \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0.36 & 0 & 0 \\ 0.18 & 0.25 & 0 \end{bmatrix}$$

A further step is now to scale all coefficient matrices by a specific scalar that is calculated for each lag matrix $\mathbf{A}_i$ by

$$a(d, s, i) = \frac{d}{i^s} \tag{A.0.1}$$

with d being the decay basis, s being the decay strength and i being the lag number.

A.1 Derivation of the estimated causal influence matrices from the p-values

A.1.1 Bivariate

If the Granger test yields the following matrices[2] containing the p-values

$$\mathbf{P}_1 = \begin{bmatrix} 0.7 & 0.02 & 0.03 \\ 0.8 & 0.5 & 0.04 \end{bmatrix}$$

$$\mathbf{P}_2 = \begin{bmatrix} 0.5 & 0.01 & 0.1 \\ 0.4 & 0.3 & 0.011 \end{bmatrix}$$

[2] Each column represents a time series, and the elements of this column represent the specific test performed. $\overline{T(k)}_t$ denotes the set of past values $\{T(k)_{t-j}, j = 1, \ldots, i\}$ of the kth time series.

Bivariate:

$$\mathbf{P}_i = \begin{bmatrix} \overline{T2}_t \Rightarrow T1_t & \overline{T1}_t \Rightarrow T2_t & \overline{T1}_t \Rightarrow T3_t \\ \overline{T3}_t \Rightarrow T1_t & \overline{T3}_t \Rightarrow T2_t & \overline{T2}_t \Rightarrow T3_t \end{bmatrix}$$

Multivariate:

$$\mathbf{P}_i = \begin{bmatrix} \overline{T2}_t \Rightarrow T1_t & \overline{T1}_t \Rightarrow T2_t & \overline{T1}_t \Rightarrow T3_t \\ \overline{T3}_t \Rightarrow T1_t & \overline{T3}_t \Rightarrow T2_t & \overline{T2}_t \Rightarrow T3_t \\ (\overline{T2}_t, \overline{T3}_t) \Rightarrow T1_t & (\overline{T1}_t, \overline{T3}_t) \Rightarrow T2_t & (\overline{T1}_t, \overline{T2}_t) \Rightarrow T3_t \end{bmatrix}$$

All on one:

$$\mathbf{P}_i = \begin{bmatrix} (\overline{T2}_t, \overline{T3}_t) \Rightarrow T1_t & (\overline{T1}_t, \overline{T3}_t) \Rightarrow T2_t & (\overline{T1}_t, \overline{T2}_t) \Rightarrow T3_t \end{bmatrix}$$

we then have to binary encode those matrices with regard to a significance level. A 0 shows that the $\mathbb{H}_0$ of Non-Granger causality could not be rejected at a predetermined significance level (in this example $\alpha = 5\%$). A 1 indicates a Granger causal influence.

$$\hat{\mathbf{C}}_1 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\hat{\mathbf{C}}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

With respect to the first lags of each time series, $\hat{\mathbf{C}}_1$ shows that

$$\begin{aligned} \overline{T1}_t &\Rightarrow T2_t \\ \overline{T1}_t &\Rightarrow T3_t \\ \overline{T2}_t &\Rightarrow T3_t \end{aligned}$$

With respect to the first two lags of each time series, $\hat{\mathbf{C}}_2$ shows that

$$\begin{aligned} \overline{T1}_t &\Rightarrow T2_t \\ \overline{T2}_t &\Rightarrow T3_t \end{aligned}$$

As we do not test if a time series Granger causes itself, which would be just an autoregressive model, we have to remove the diagonal of all $\mathbf{C}_i^{init}$ matrices. Furthermore, we have to transpose those matrices to being able to compare the estimated causal influences $\hat{\mathbf{C}}_i$ with the true causal influences $\mathbf{C}_i$[3]. Each column now represent one time series.

Comparing the estimated causal influences $\hat{\mathbf{C}}_1$ with the true causal influences $\mathbf{C}_1$ is unproblematic, because it we only consider the results of a Granger test that use the first lag terms. All subsequent matrices $\hat{\mathbf{C}}_i$ with $i > 1$ must be slightly transformed because the Granger

[3] Removing the main diagonal leads to

$$\mathbf{C}_{*1} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \quad and \quad \mathbf{C}_{*2} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

Transposing those leads to

$$\mathbf{C}_1 := \mathbf{C}_{*1}^T = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad and \quad \mathbf{C}_2 := \mathbf{C}_{*2}^T = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

test uses more lag terms, but the number of derived test statistics does not change. This results in the problem that we have only one matrix with p-values[4], but more than one matrix containing the true causal influences and coefficients for the lag terms under consideration.

A solution is to repeat the estimated causal influences matrix. In our example with two lags we only have to adjust the second one. Hence we duplicate $\hat{\mathbf{C}}_2$ to derive the aggregated matrix/array $\hat{\mathbf{C}}_2^A$ with

$$\hat{\mathbf{C}}_2^A = \left[\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \right]$$

The first matrix stays as it is

$$\hat{\mathbf{C}}_1^A = \hat{\mathbf{C}}_1 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

A.1.2 All on one (multivariate)

The $\mathbf{P}_i$ matrices that contain the p-values of the Granger tests are give as an example by:

$$\mathbf{P}_1 = \begin{bmatrix} 0.6 & 0.04 & 0.01 \end{bmatrix}$$

$$\mathbf{P}_2 = \begin{bmatrix} 0.76 & 0.069 & 0.001 \end{bmatrix}$$

The resulting $\hat{\mathbf{C}}_i$ with $\alpha = 5\%$ are

$$\hat{\mathbf{C}}_1 = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

$$\hat{\mathbf{C}}_2 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

[4] And hence one estimated causal influence matrix.

With respect to the first lags of each time series, $\hat{\mathbf{C}}_1$ shows that

$$(\overline{T1}_t, \overline{T3}_t) \Rightarrow T2_t$$
$$(\overline{T1}_t, \overline{T2}_t) \Rightarrow T3_t$$

With respect to the first two lags of each time series, $\hat{\mathbf{C}}_2$ shows that

$$(\overline{T1}_t, \overline{T2}_t) \Rightarrow \quad T3_t$$

Again, we have the dimensionality problem that the shape of $\mathbf{C}_i$ is not equal to $\hat{\mathbf{C}}_i$. Therefore, $\hat{\mathbf{C}}_i$ must be transformed in such a way that the causal influence of each time series alone is visible[5]. Hence, $\hat{\mathbf{C}}_1$ is transformed to

$$\hat{\mathbf{C}}_1 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

by just repeating/duplicating the first column.

To compare the estimated causal influences $\hat{\mathbf{C}}_i$ with the true causal influences $\mathbf{C}_i$ for $i > 1$, one must aggregate in the same way as in the bivariate case. Hence for $i = 2$ we derive

$$\hat{\mathbf{C}}_2^A = \left[\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \right]$$

[5] If $(\overline{T1}_t, \overline{T3}_t) \Rightarrow T2_t$, it is concluded that also $\overline{T1}_t \Rightarrow T2_t$ and $\overline{T3}_t \Rightarrow T2_t$ hold. $\overline{T1}_t$ and $\overline{T3}_t$ together and individually Granger cause $T2_t$.

A.2 Correlation

Two correlation coefficients are used:

1. The Pearson correlation coefficient (Virtanen et al., 2020, pearsonr) to test whether a high true coefficient in $\mathbf{A}_i$ leads to an estimated causal influence, which is estimated using the p-values $\mathbf{P}_i$. Optimally, a low and thus significant p-value should be accompanied by a high true coefficient in $\mathbf{A}_i$.
2. The point biserial correlation coefficient[6] (Virtanen et al., 2020, pointbiserialr) tests wether low p-values $\mathbf{P}_i$ and a true causal influence $\mathbf{C}_i$ correlate.

A.2.1 Bivariate

We face the same problem as before, that we only have on p-value matrix but several true causal influence and coefficient matrices when considering more than one lag. Thence, the p-value matrix is transformed in similar fashion

$$\hat{\mathbf{P}}_1^A = \hat{\mathbf{P}}_1 = \begin{bmatrix} 0.7 & 0.02 & 0.03 \\ 0.8 & 0.5 & 0.04 \end{bmatrix}$$

For the first lag nothing changes, but for the second lag we repeat/duplicate the p-value matrix

$$\hat{\mathbf{P}}_2^A = \left[\begin{bmatrix} 0.5 & 0.01 & 0.1 \\ 0.4 & 0.3 & 0.011 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.01 & 0.1 \\ 0.4 & 0.3 & 0.011 \end{bmatrix} \right]$$

As for the Pearson correlation, we compare the true coefficient matrices to the transformed p-value matrices. To compare both type, we have to transform the true coefficient matrices by removing the main diagonal to the apply transposition. Afterwards, we aggregate those matrices like in the following

$$\mathbf{A}{*}_1^A = \mathbf{A}{*}_1 = \begin{bmatrix} 0 & 0.40 & 0.30 \\ 0 & 0 & 0.37 \end{bmatrix}$$

$$\mathbf{A}{*}_2^A = \left[\begin{bmatrix} 0 & 0.40 & 0.30 \\ 0 & 0 & 0.37 \end{bmatrix}, \begin{bmatrix} 0 & 0.36 & 0.18 \\ 0 & 0 & 0.25 \end{bmatrix} \right]$$

[6] Is used because $\mathbf{C}_i$ contains binary variables, and $\mathbf{P}_i$ contains continuous variables.

The point biserial correlation takes the transformed true causal influence matrices and the transformed estimated causal influence matrices[7] as input.

$$\mathbf{C}_1^A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{C}_2^A = \left[\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \right]$$

To derive the correlation performance of a Granger test with two lags, we take $\hat{\mathbf{P}}_2^A$ and test the correlation with $\mathbf{C}_2^A$ (point biserial) or $\mathbf{A}{*}_2^A$ (Pearson).

A.2.2 All on one (multivariate)

The logic stays the same, we first duplicate/repeat the first row to then duplicate/repeat this matrix

$$\hat{\mathbf{P}}_1^A = \hat{\mathbf{P}}_1 = \begin{bmatrix} 0.6 & 0.04 & 0.01 \\ 0.6 & 0.04 & 0.01 \end{bmatrix}$$

$$\hat{\mathbf{P}}_2^A = \left[\begin{bmatrix} 0.76 & 0.069 & 0.001 \\ 0.76 & 0.069 & 0.001 \end{bmatrix}, \begin{bmatrix} 0.76 & 0.069 & 0.001 \\ 0.76 & 0.069 & 0.001 \end{bmatrix} \right]$$

These matrices now have the same form and dimensionality as $\mathbf{C}_2^A$ and $\mathbf{A}{*}_2^A$ to derive the point biserial or Pearson correlation coefficient.

Complete multivariate When conducting $K \times 2^{K-1} - 1$ different Granger tests, it is not possible to approximate the p-value and causal influence of a single time series on another one because of lacking transitivity.

As an example: If $(\overline{T1}_t, \overline{T2}_t, \overline{T3}_t) \Rightarrow T4_t$, $(\overline{T1}_t, \overline{T2}_t) \Rightarrow T4_t$ and $(\overline{T1}_t, \overline{T3}_t) \Rightarrow T4_t$, but $(\overline{T2}_t, \overline{T3}_t) \not\Rightarrow T4_t$. How do we now approximate the causal influence of each single times series $\overline{T1}_t$, $\overline{T2}_t$ and $\overline{T3}_t$ on $T4_t$!? Because this is not feasible, we cannot perform correlation analyses.

[7] Replace every element in $\hat{\mathbf{P}}_1^A$ and $\hat{\mathbf{P}}_2^A$ that is lager than the significance level with a 0 and otherwise with a 1

A.3 Accuracy rate

Analyzing the accuracy of each Granger test with regard to the detected and thus estimated causal influences incorporates a comparison of the arrays $\mathbf{C}_i^A$ and $\hat{\mathbf{C}}_i^A$. This is achieved through setting up a confusion matrix $\mathbf{CM}_i$ for each lag i. The confusion matrix $\mathbf{CM}_i$ of lag i summarizes $i \times (K-1) \times K$ comparison results. The structure of these confusion matrices is shown in the Figure A.1 as an example.

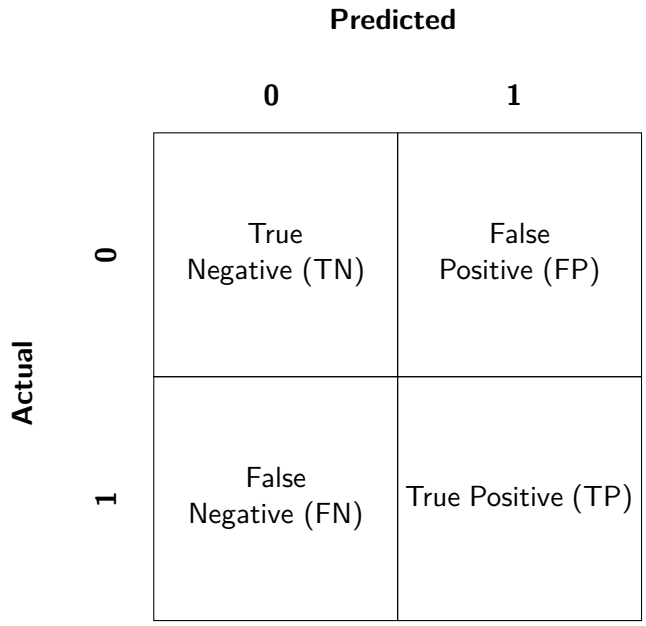


Figure A.1: Confusion matrix with 0 for non-causality and 1 for causality

To calculate the prediction accuracy, one simply divides the trace of $\mathbf{CM}_i$ by the sum of $\mathbf{CM}_i$.

Complete multivariate Again, the problem of approximation arises because it is not possible to construct a confusion matrix and derive the accuracy rate from it since we cannot convert the estimated causal influence matrix into a form that matches that of the true causal influence matrix.

As a workaround, we calculate the row-wise average of the estimated and the true causal influence matrix. Thus, per time series, we can now compare the percentage of causal influences with the percentage of estimated causal influences.

A.4 Drawing coefficients from the Beta distribution

The coefficient matrices are systematically drawn from the percent point function of the Beta distribution (see Figure A.3) to obtain values $v_{(\cdot)} \in [0, 1]$[8].

The probability density function[9] is calculated by

$$PDF : f(x, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}$$

and illustrated for different α and β values in Figure A.2.

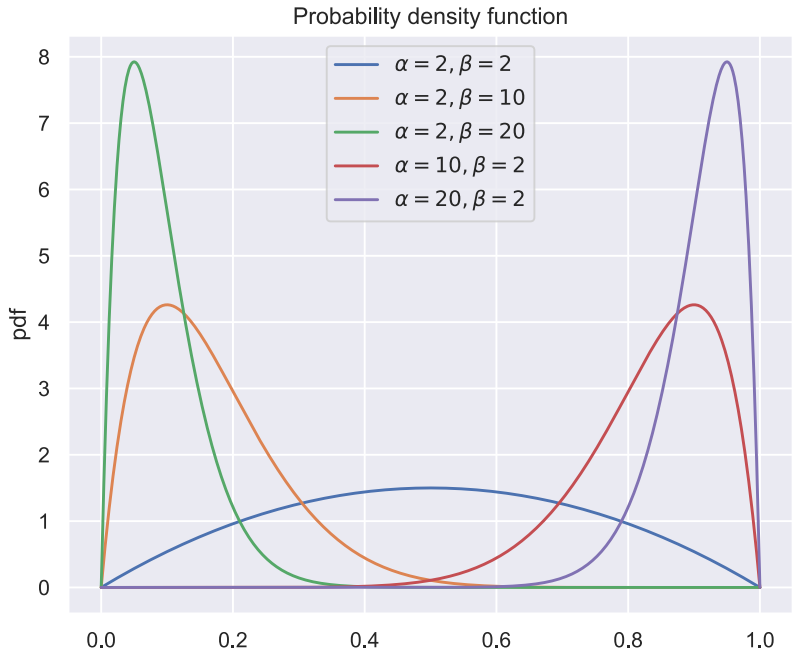


Figure A.2: Probability density function of the Beta distribution

The first step is to generate the coefficients matrix $\boldsymbol{A}_i$ by drawing random diagonal elements that represent the autoregressive coefficients. Example with $k = 1, \ldots, 4$ and $p = 1$:

[8] The coefficients are in the range between zero and one to prevent the time series from exploding. The setting `allow_neg_coeffs=True` in `TS_Sim` allows for negative coefficients as well.

[9] For details see (Virtanen et al., 2020, scipy.stats.beta).

$$\mathbf{A}_1^{AR} = \begin{bmatrix} 0.8666 & 0 & 0 & 0 \\ 0 & 0.2631 & 0 & 0 \\ 0 & 0 & 0.1314 & 0 \\ 0 & 0 & 0 & 0.0416 \end{bmatrix}$$

In a next step we proceed line by line and calculate the probability that $X \leq x$ using the autoregressive coefficient as inputs of the cdf[10].

Take the third row as an example: For 0.1314 we derive a probability of $\mathcal{P}(X \leq 0.1314) = 0.0476$ (see Listing 1). We then look at how many elements are to the left and right of each diagonal element. For the third row, we find two elements on the left and one on the right.

Now we divide 0.0476 by the enumerated numbers one to the number of rows on the left side[11]. This gives the probabilities 0.01576 and 0.0315 The same is done for the right side. We find only one element, so we divide 0.0476 by two and get 0.0236. Using the inverse of the cumulative distribution function, we obtain the respective quantiles that represent the respective new coefficients. Applying this logic leads to the following coefficient matrix

$$\mathbf{A}_1 = \begin{bmatrix} 0.8666 & 0.6466 & 0.4838 & 0.3171 \\ 0.1801 & 0.2631 & 0.2104 & 0.1452 \\ \mathbf{0.0743} & \mathbf{0.1063} & 0.1314 & \mathbf{0.0916} \\ 0.0206 & 0.0293 & 0.036 & 0.0416 \end{bmatrix}$$

[10] The cdf and ppf have $\alpha = 2$ and $\beta = 2$.

[11] With two elements we divide 0.0476 by two and three because dividing 0.0476 by one gives 0.0476. Thus we start enumerating from two to the number of elements, but adding one.

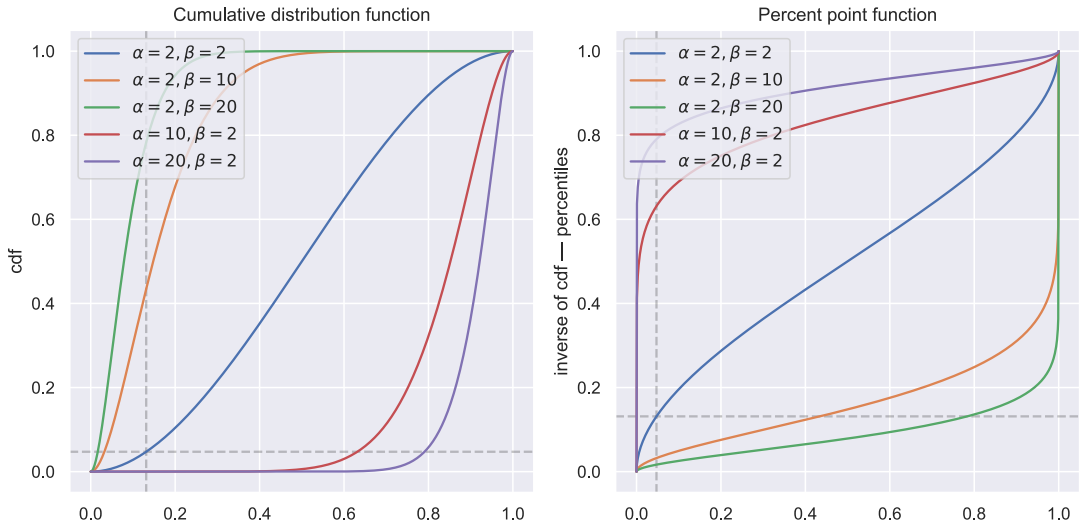


Figure A.3: Beta distribution (cdf and ppf)

Figure A.3 shows the cumulative distribution function and the percentage point function of the Beta distribution, with the values `cdf` and `inv_cdf` from Listing 1 highlighted by dashed grey lines.

```
import scipy.stats as scp
import numpy as np

np.random.seed(128)
ar = np.abs(np.random.random(4)) # ar[2] = 0.131408478173487
beta = scp.beta(a=2, b=2)
cdf = beta.cdf(x=ar[2]) # cdf = 0.047266191760118675
inv_cdf = beta.ppf(q=cdf) # inv_cdf = 0.131408478173487
```

Listing 1: Using `cdf()` and `ppf()` of `scipy.stats.beta`

Scaling up alpha

With $\alpha = 10 \wedge \beta = 2$:

$$\mathbf{A}_1 = \begin{bmatrix} 0.8666 & 0.8421 & 0.8086 & 0.7545 \\ 0.2455 & 0.2631 & 0.2527 & 0.2358 \\ 0.1177 & 0.1262 & 0.1314 & 0.1226 \\ 0.0362 & 0.0388 & 0.0404 & 0.0416 \end{bmatrix}$$

With $\alpha = 20 \wedge \beta = 2$:

$$\mathbf{A}_1 = \begin{bmatrix} 0.8666 & 0.8543 & 0.8371 & 0.8086 \\ 0.2542 & 0.2631 & 0.2579 & 0.2491 \\ 0.1244 & 0.1288 & 0.1314 & 0.1269 \\ 0.0388 & 0.0402 & 0.0410 & 0.0416 \end{bmatrix}$$

Except for the diagonal elements, the coefficients become more and more similar line by line, approaching the respective diagonal value.

Scaling up beta

With $\alpha = 2 \wedge \beta = 10$:

$$\mathbf{A}_1 = \begin{bmatrix} 0.8666 & 0.2266 & 0.1480 & 0.0876 \\ 0.1265 & 0.2631 & 0.1620 & 0.0938 \\ 0.0617 & 0.0969 & 0.1314 & 0.0798 \\ 0.0195 & 0.0283 & 0.0354 & 0.0416 \end{bmatrix}$$

With $\alpha = 2 \wedge \beta = 20$:

$$\mathbf{A}_1 = \begin{bmatrix} 0.8666 & 0.1232 & 0.0786 & 0.0458 \\ 0.0776 & 0.2631 & 0.1040 & 0.0558 \\ 0.0472 & 0.0819 & 0.1314 & 0.0639 \\ 0.0180 & 0.0269 & 0.0345 & 0.0416 \end{bmatrix}$$

Apart from the diagonal elements, the coefficients become more and more similar line by line and approach zero.

B Partial Granger causality test

Sticking to the notation of the bivariate setting with X_t and Y_t, one can test if $\overline{\overline{X}}_{t-i} \Rightarrow Y_t$, where $\overline{\overline{X}}_{t-i}$ denotes a set containing just one variable, namely the ith lagged variable X_t[1]. A usual F-test is then employed to decide between an unrestricted and a restricted model.

The restricted model regresses Y_t on $\overline{Y}_t$ to obtain the residual sum of squares RSS_R. Thus, this is a normal autoregressive model regressing Y_t on its i lagged values.

The unrestricted model then derives the residual sum of squares RSS_{UR} by regressing Y_t on all i autoregressive terms and the ith lagged value of X_t[2].

The null hypothesis $\mathbb{H}_0$ states that the coefficient of the ith lagged value of X_t does not belong in the regression and thereby do not Granger cause Y_t.

$$F = \frac{(RSS_R - RSS_{UR})/\#\overline{\overline{X}}_{t-i}}{RSS_{UR}/(n-k)}$$

$\#\overline{\overline{X}}_{t-i}$ denotes the cardinality of $\overline{\overline{X}}_{t-i}$, that is one is this case, and $(n-k)$ is the difference between the number of samples and the number of coefficients estimated in the unrestricted regression (see Gujarati and Porter, 2009, chapter 17, p. 653 et seq.).

Note: This concept does not eliminate the influence of additional variables as does the concept of partial Granger causality proposed by Guo et al. (Guo et al., 2008). In this paper, the term "partial" is used to clarify that the Granger causality test is not aggregated and therefore uses only one specific lag coefficient.

[1] Using the lag operator this can be written as $L^i X_t \Rightarrow Y_t$ with $L^i X_t = X_{t-i}$.

[2] A concrete example testing wether the 5th lag of X_t Granger causes Y_t looks like:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \beta_5 Y_{t-5} + u_t \qquad \text{(Restricted model)}$$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \beta_5 Y_{t-5} + \alpha_5 X_{t-5} + u_t \qquad \text{(Unrestricted model)}$$

C Explanation of the diagrams of the sensitivity analysis

C.1 Accuracy rate

If one iterates over a parameter of the TS_S and still considers several lags, this leads to a three-dimensional representation. Since static 3D plots are difficult to interpret on paper, the dimension of the lags (formerly abscissa in the single analysis) is summarized.

The plot for accuracy per package thus looks like Figure C.1. The blue line is the reference development of the specified lag (in this case lag 7[1]). All other elements are derived from the summary statistics using the results of all other lags. This gives you a first impression of the total range.

C.2 Correlation

The correlation diagram can be interpreted similarly to the individual analysis. Compared to the accuracy plot above, the results for each lag are not summarized but represented by individual lines. If a correlation coefficient is significant, this is indicated by a solid and colored line/marker and explicitly stated in the legend. An example is shown in Figure C.2.

[1] 1:7 in the plot only indicates that we consider the aggregated values

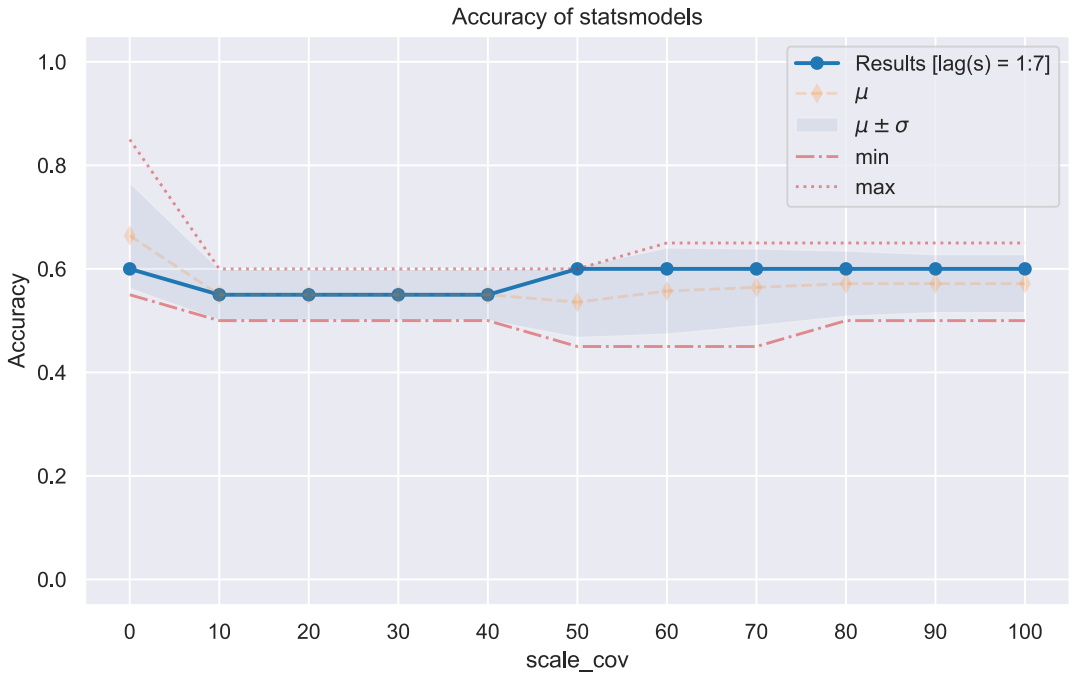


Figure C.1: Sensitivity analysis: Accuracy when scaling up the correlation between the time series

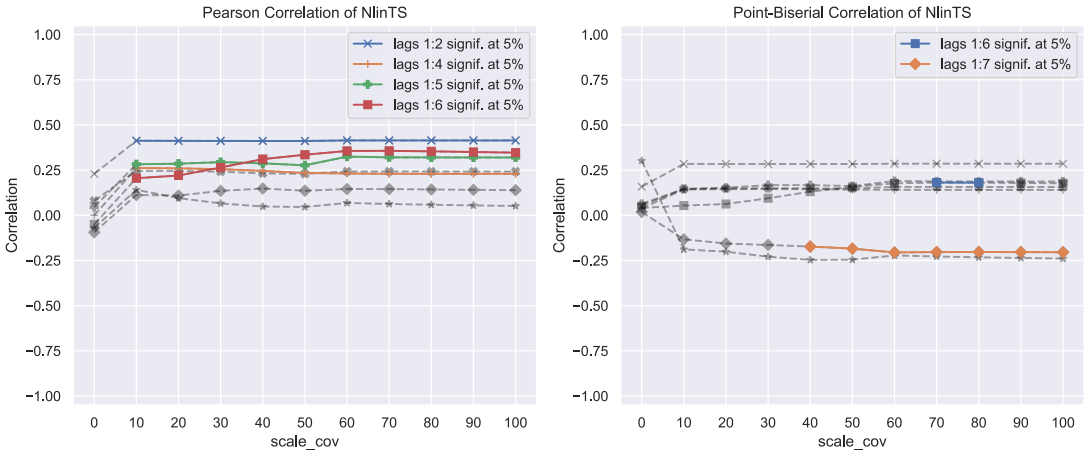


Figure C.2: Sensitivity analysis: Correlation coefficients when scaling up the correlation between the time series

D Figures

D.1 Time series

D.1.1 With structural breaks

See Figure D.1. The covariance matrix of the error terms specific to each breakpoint[1] is given by:

```
error covariance tensor:
[[[7.4822 0.     0.     0.     0.    ]
  [0.     3.2344 0.     0.     0.    ]
  [0.     0.     0.0163 0.     0.    ]
  [0.     0.     0.     2.2151 0.    ]
  [0.     0.     0.     0.     3.9238]]

 [[0.2725 0.     0.     0.     0.    ]
  [0.     0.6194 0.     0.     0.    ]
  [0.     0.     0.0002 0.     0.    ]
  [0.     0.     0.     0.0735 0.    ]
  [0.     0.     0.     0.     0.4701]]

 [[1.0898 0.     0.     0.     0.    ]
  [0.     0.3247 0.     0.     0.    ]
  [0.     0.     4.9086 0.     0.    ]
  [0.     0.     0.     2.4363 0.    ]
  [0.     0.     0.     0.     0.9596]]

 [[0.1722 0.     0.     0.     0.    ]
  [0.     0.9773 0.     0.     0.    ]
  [0.     0.     0.0229 0.     0.    ]
  [0.     0.     0.     0.2954 0.    ]
  [0.     0.     0.     0.     0.1201]]]
```

[1] With $n_{sb} = 3$ breakpoints we have four sections. Consequently, four covariance matrices are displayed. When a time series has no breakpoints, like the second time series, the first covariance matrix is used to generate its error terms over the entire time domain.

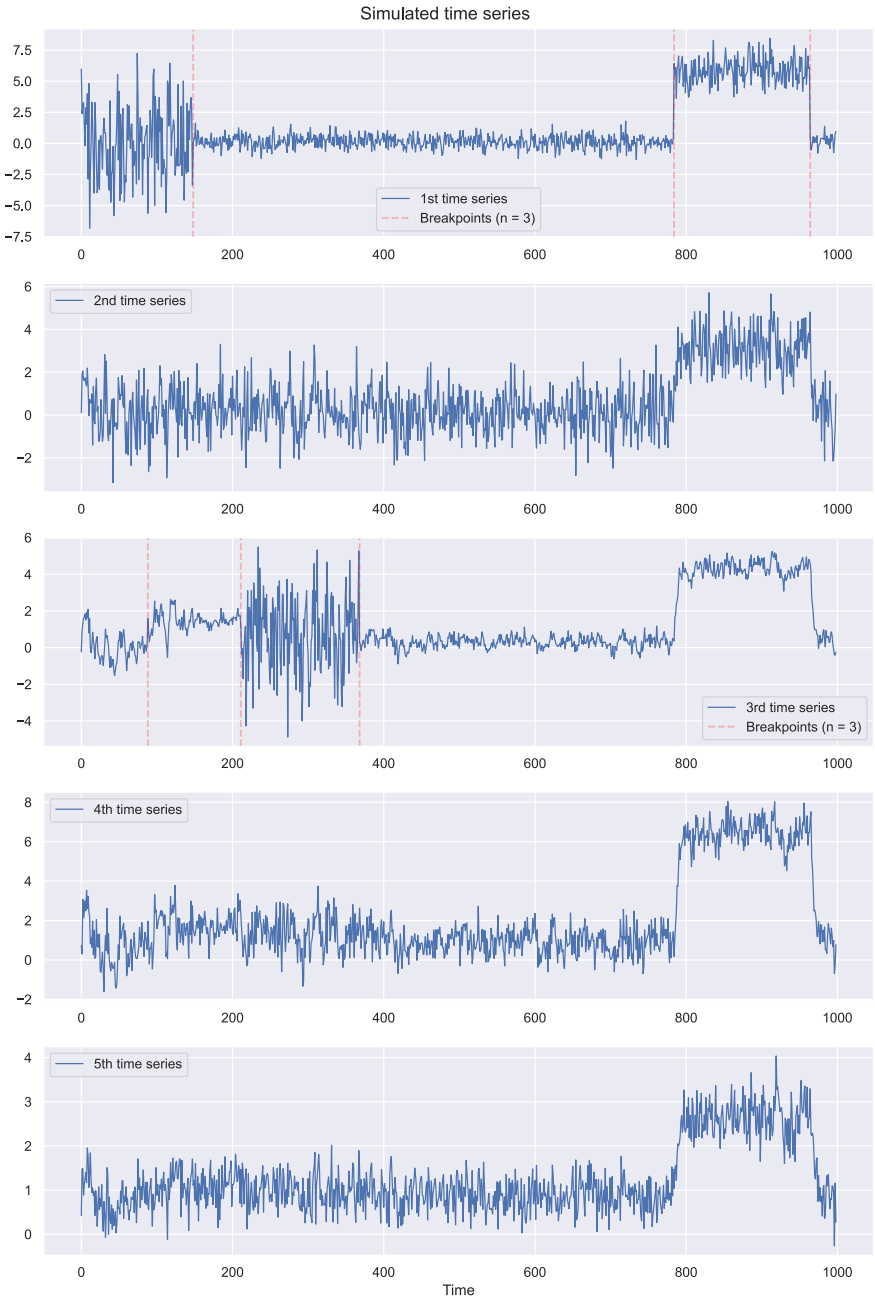


Figure D.1: Five simulated time series with structural breaks

D.1.2 With a time trend

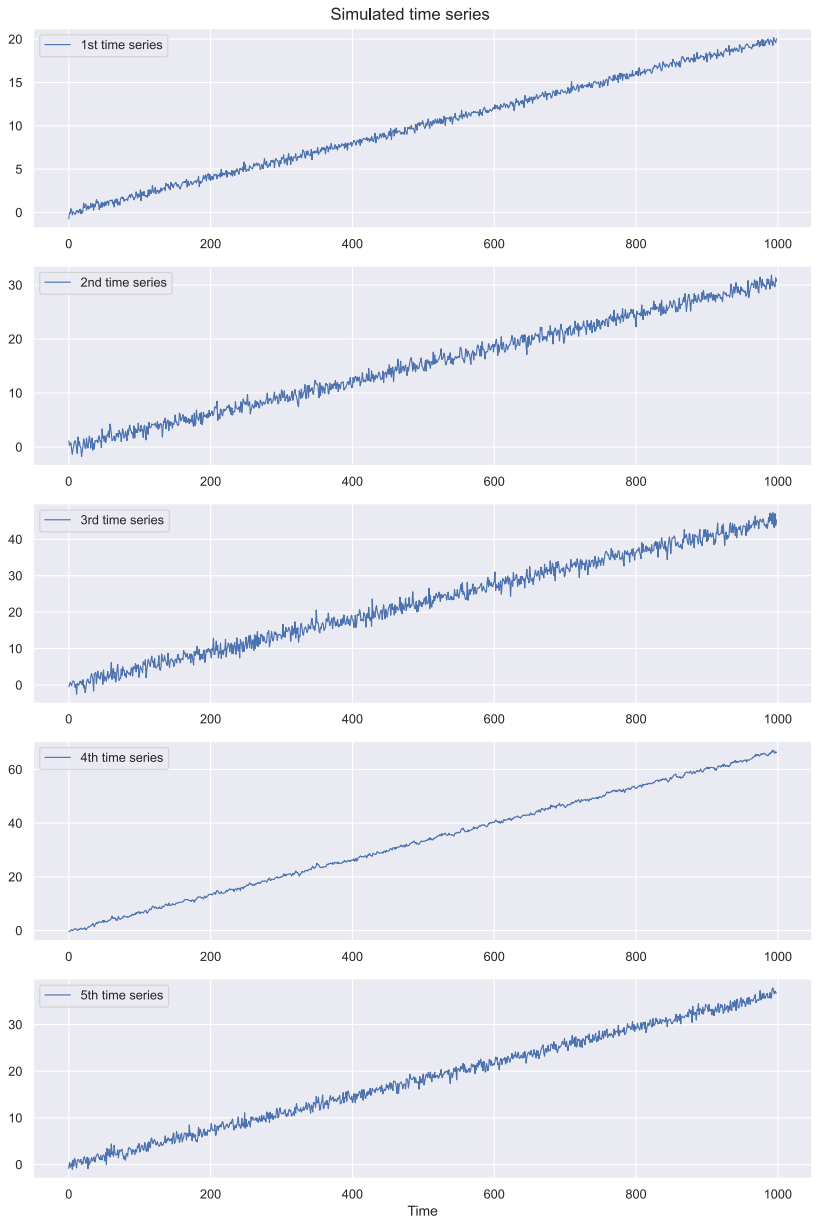
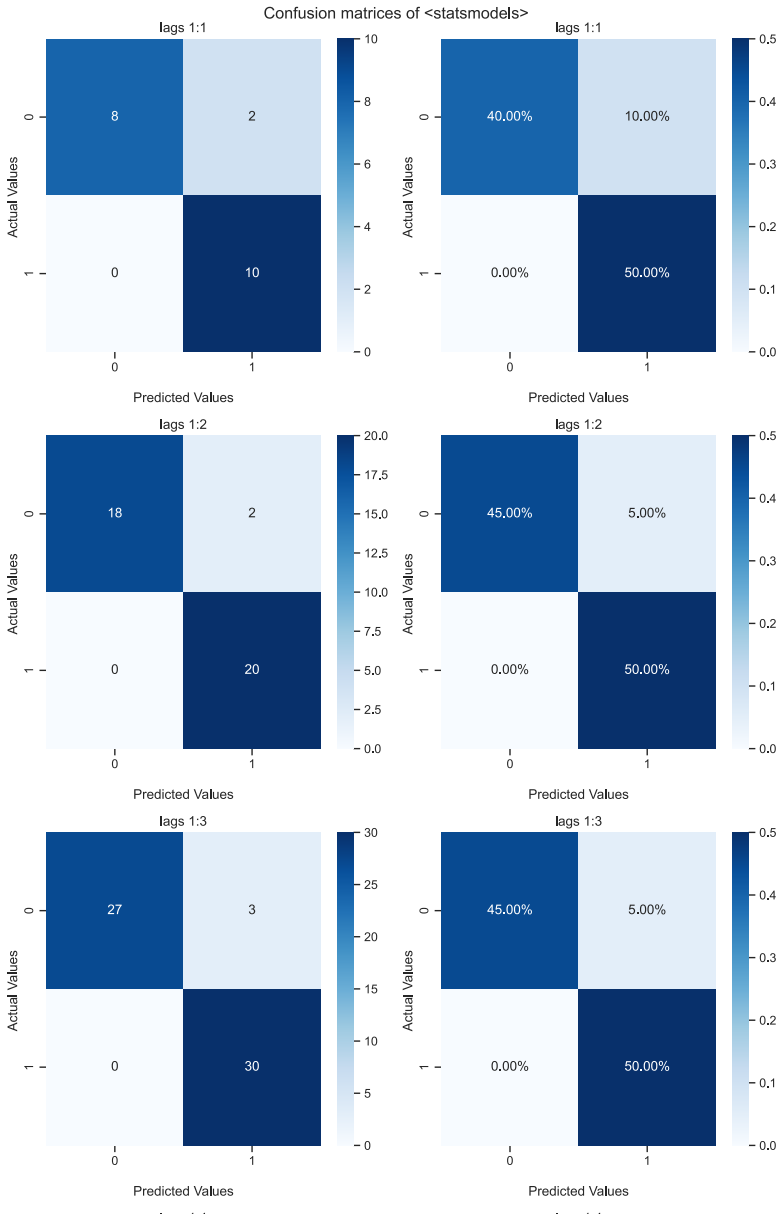


Figure D.2: Five simulated time series a time trend

D.2 Single Analysis

D.2.1 Confusion matrices of statsmodels (bivariate)



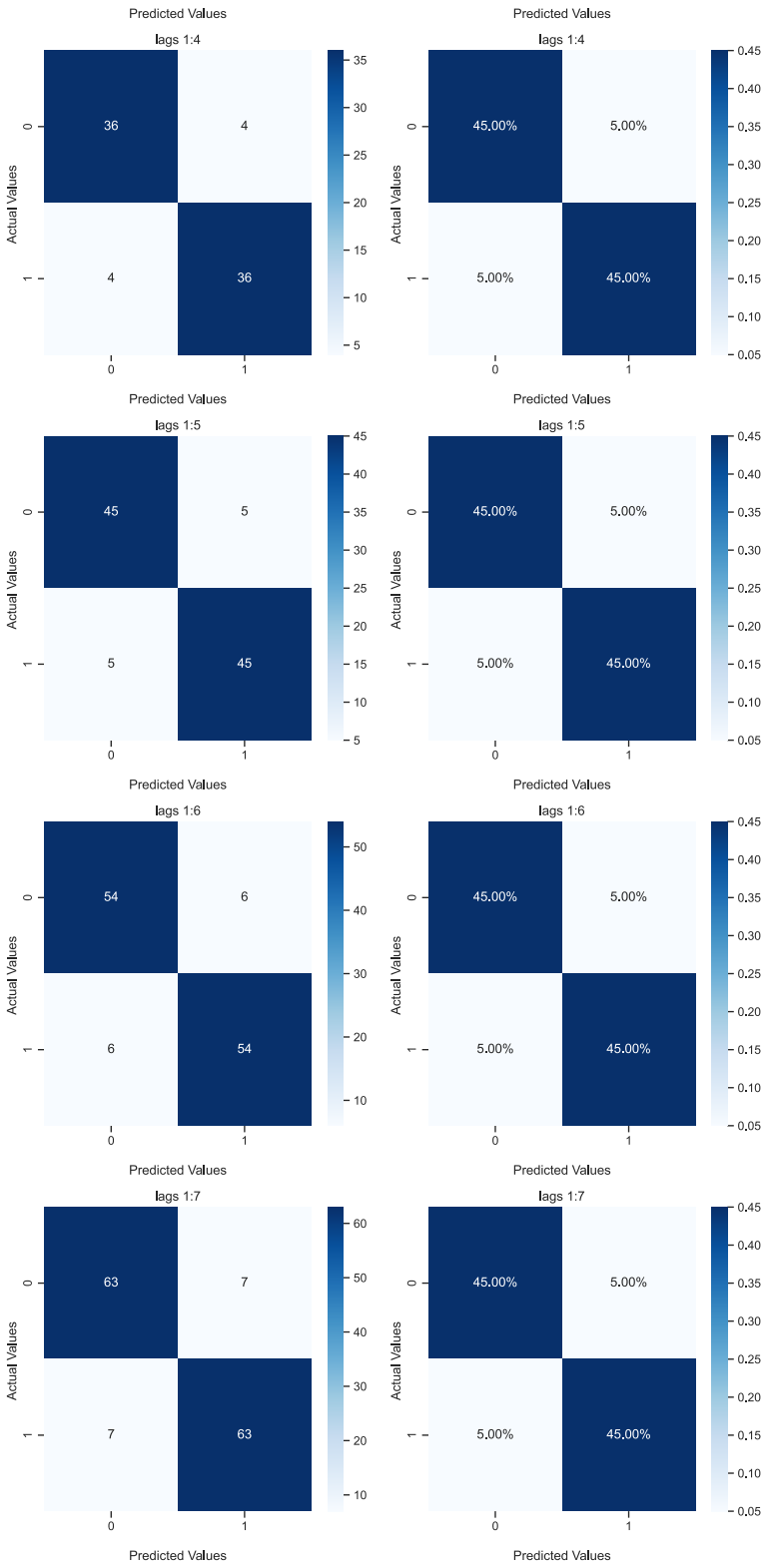
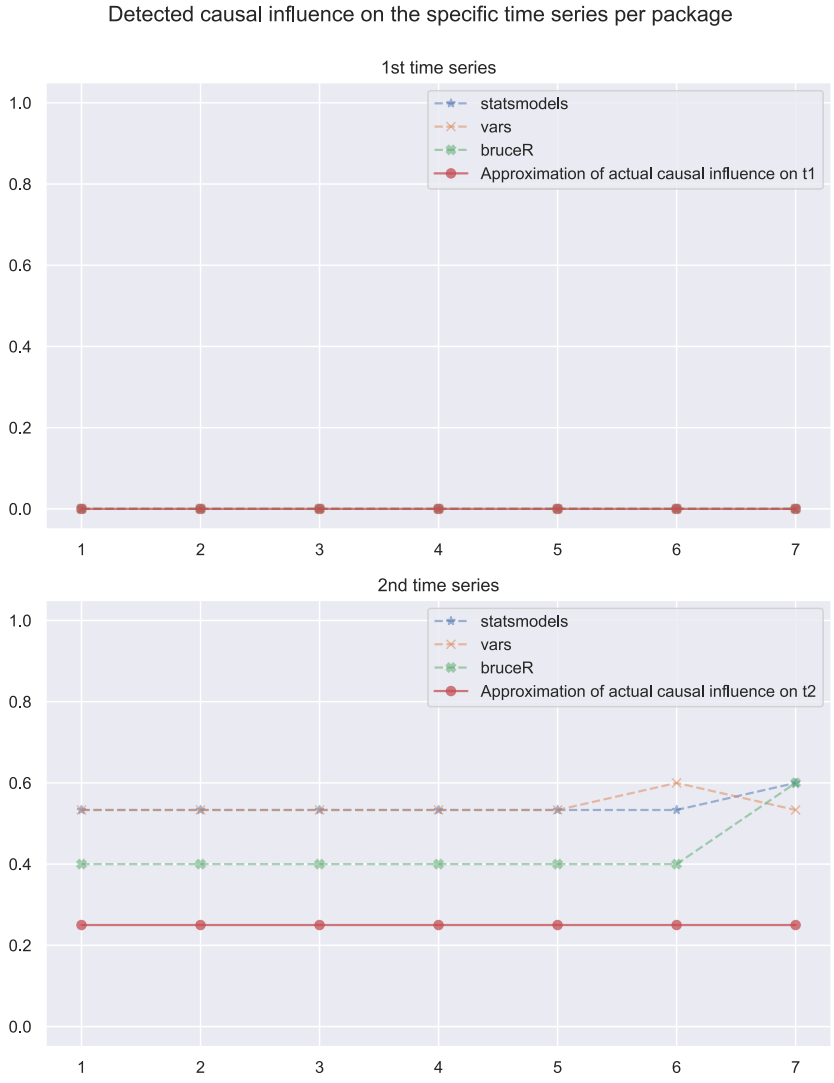


Figure D.3: All confusion matrices by lags of `statsmodels`

D.2.2 True multivariate comparison



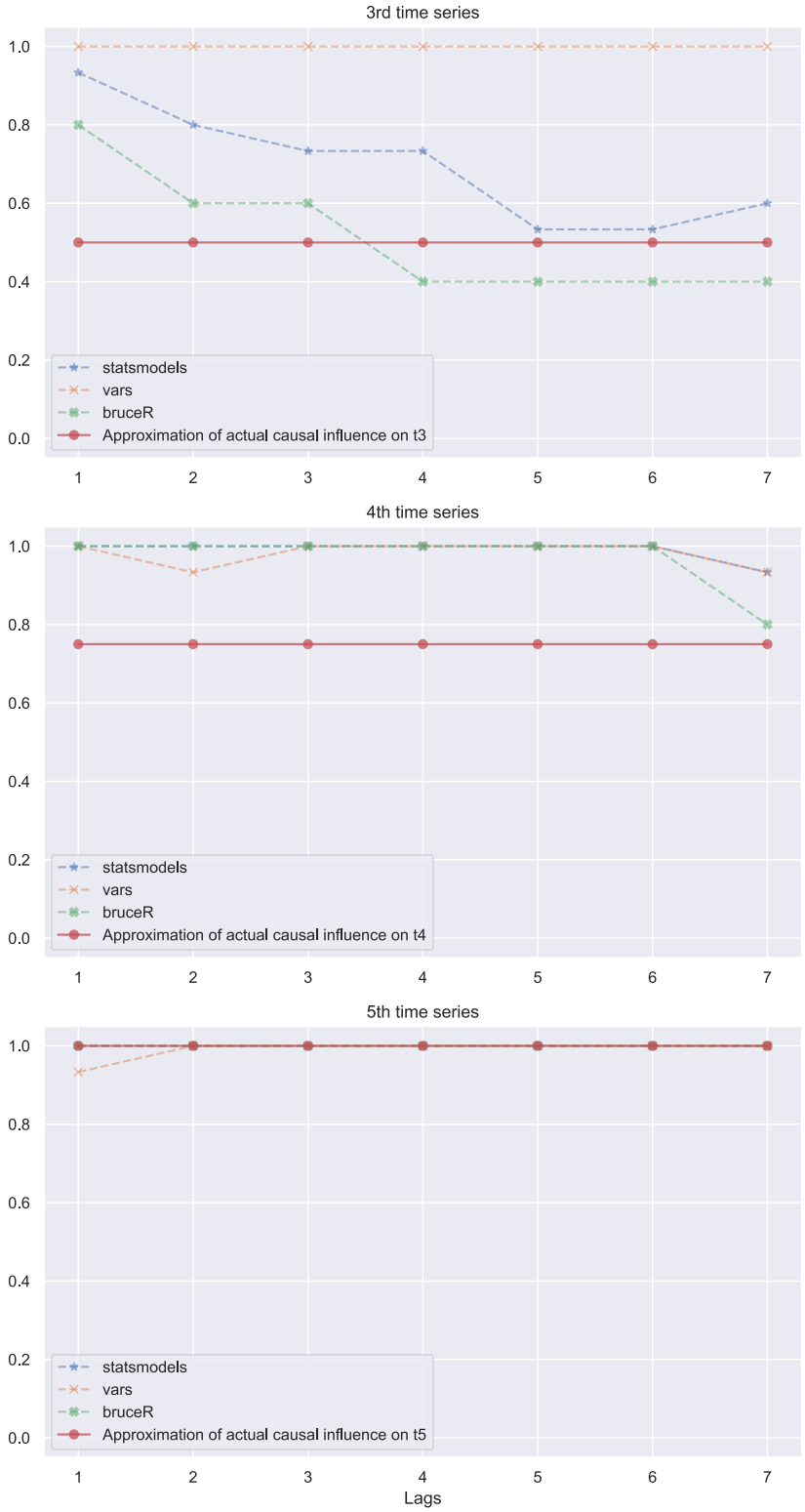
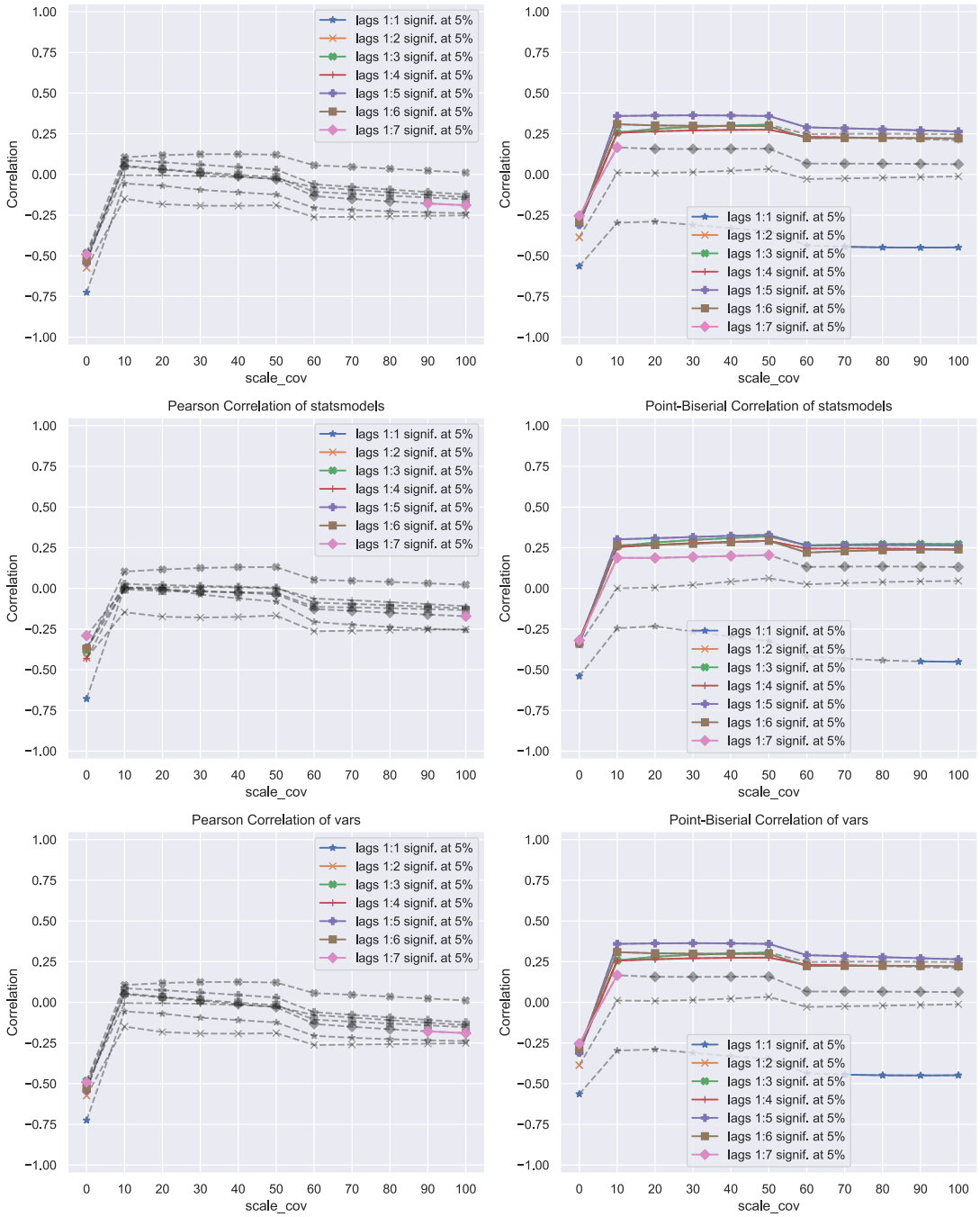


Figure D.4: Approximated accuracy per time series in the *true multivariate* case

D.2.3 Sensitivity analysis

D.2.3.1 Bivariate



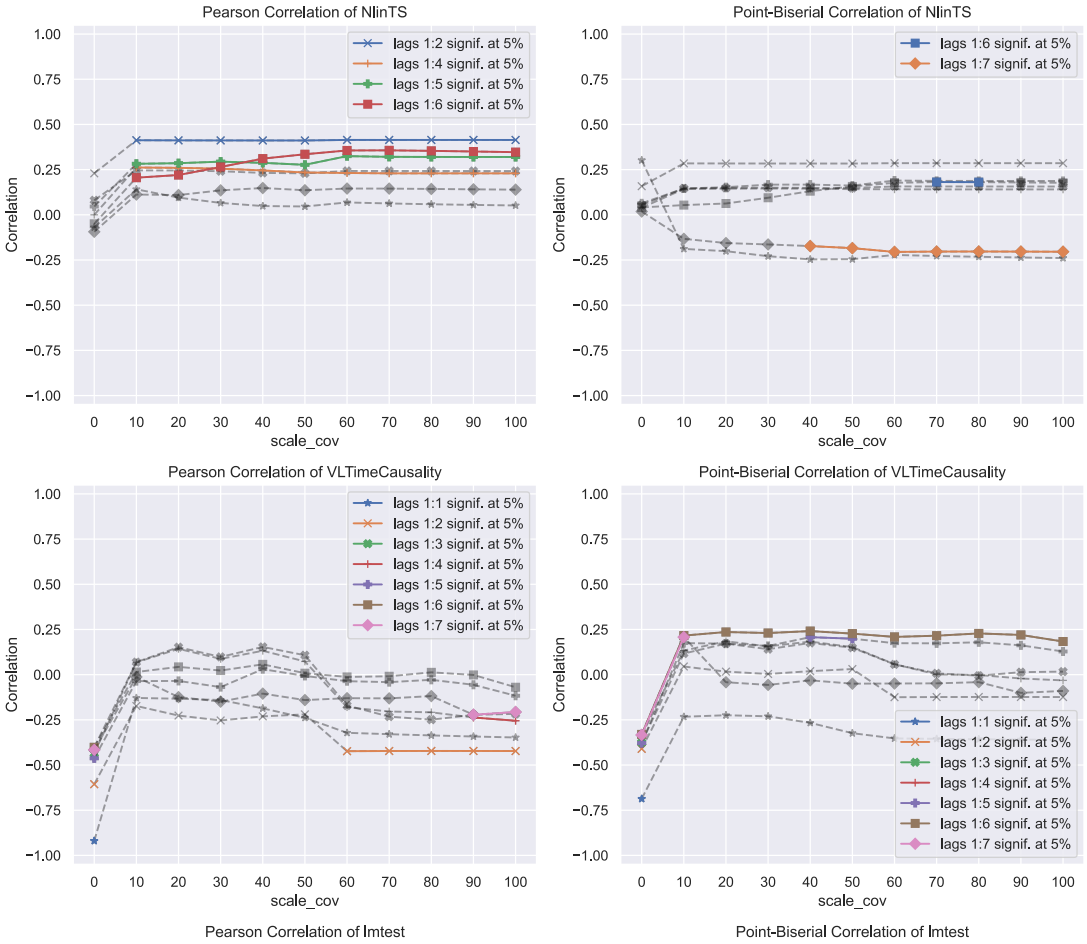
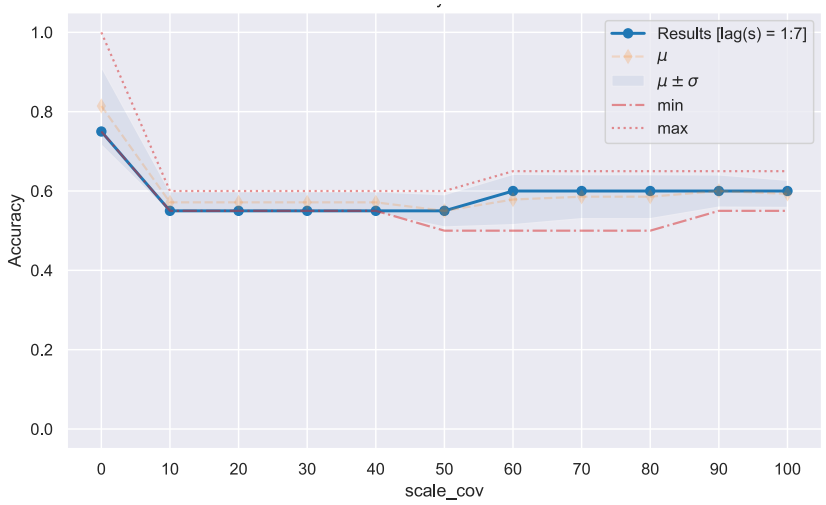
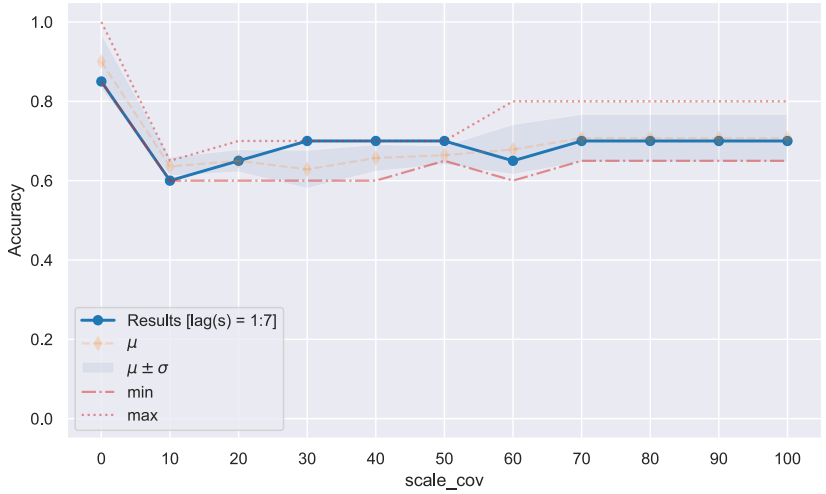


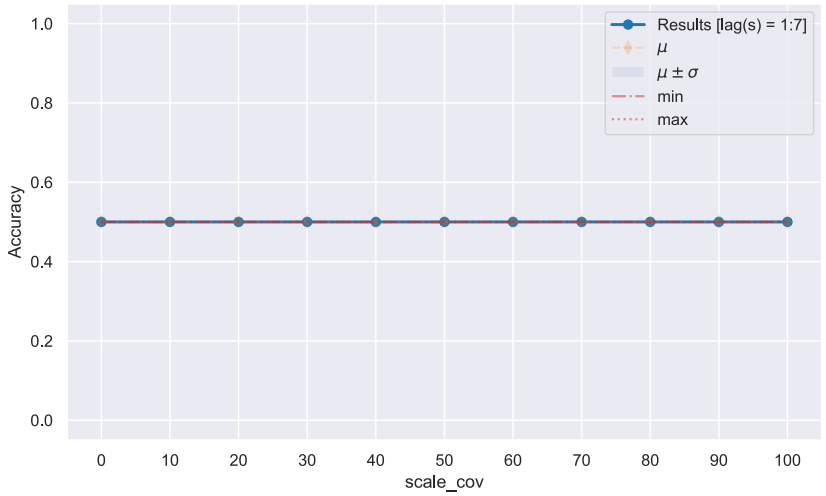
Figure D.5: Pearson correlation plots (bivariate | sensitivity analysis | covariance upscaling)



Accuracy of VLTimeCausality



Accuracy of NlinTS



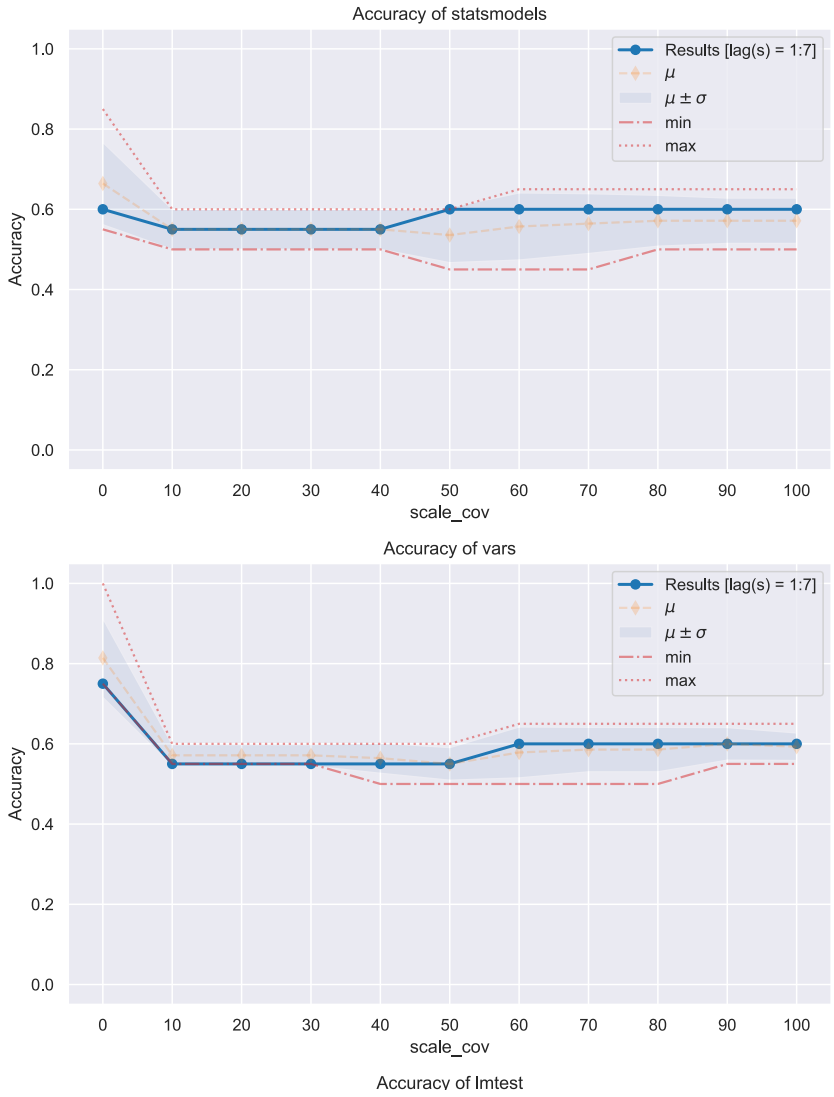


Figure D.6: Accuracy plots (bivariate | sensitivity analysis | covariance upscaling)

D.2.3.2 All on one

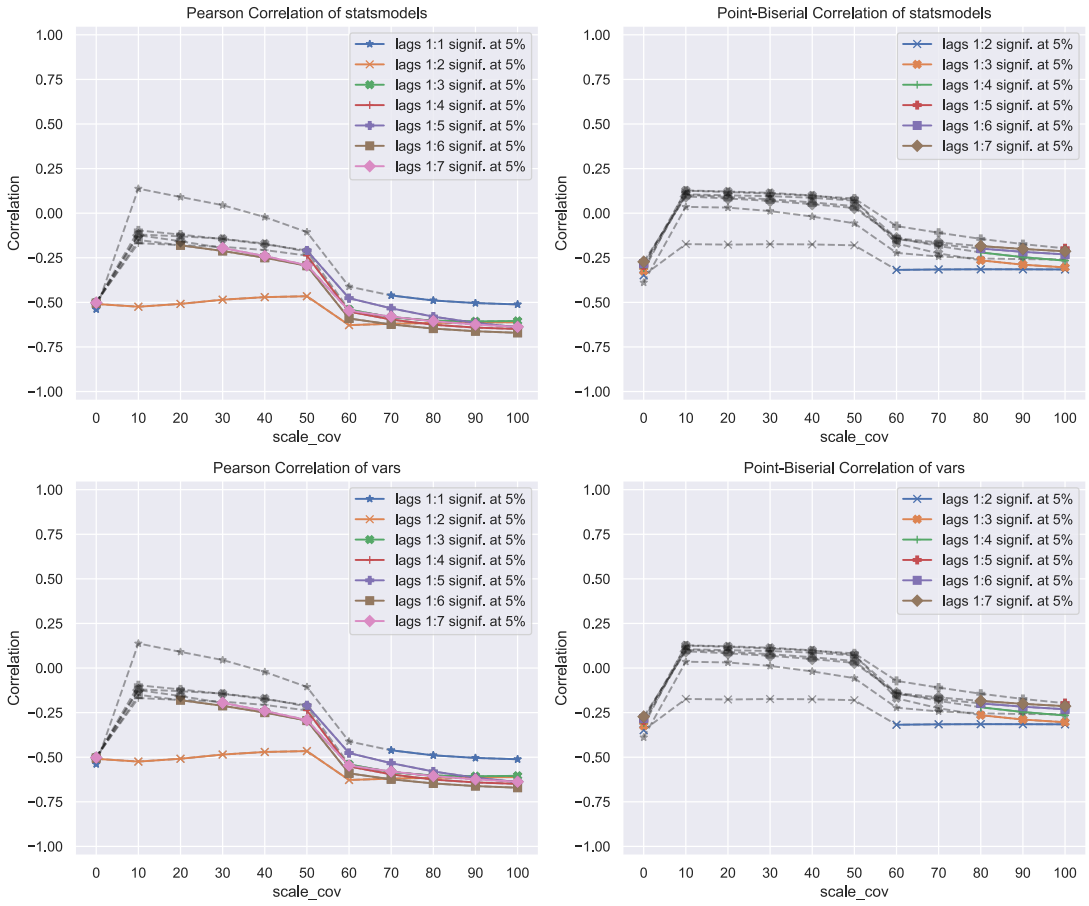


Figure D.7: Pearson correlation plots (all on one | sensitivity analysis | covariance upscaling)

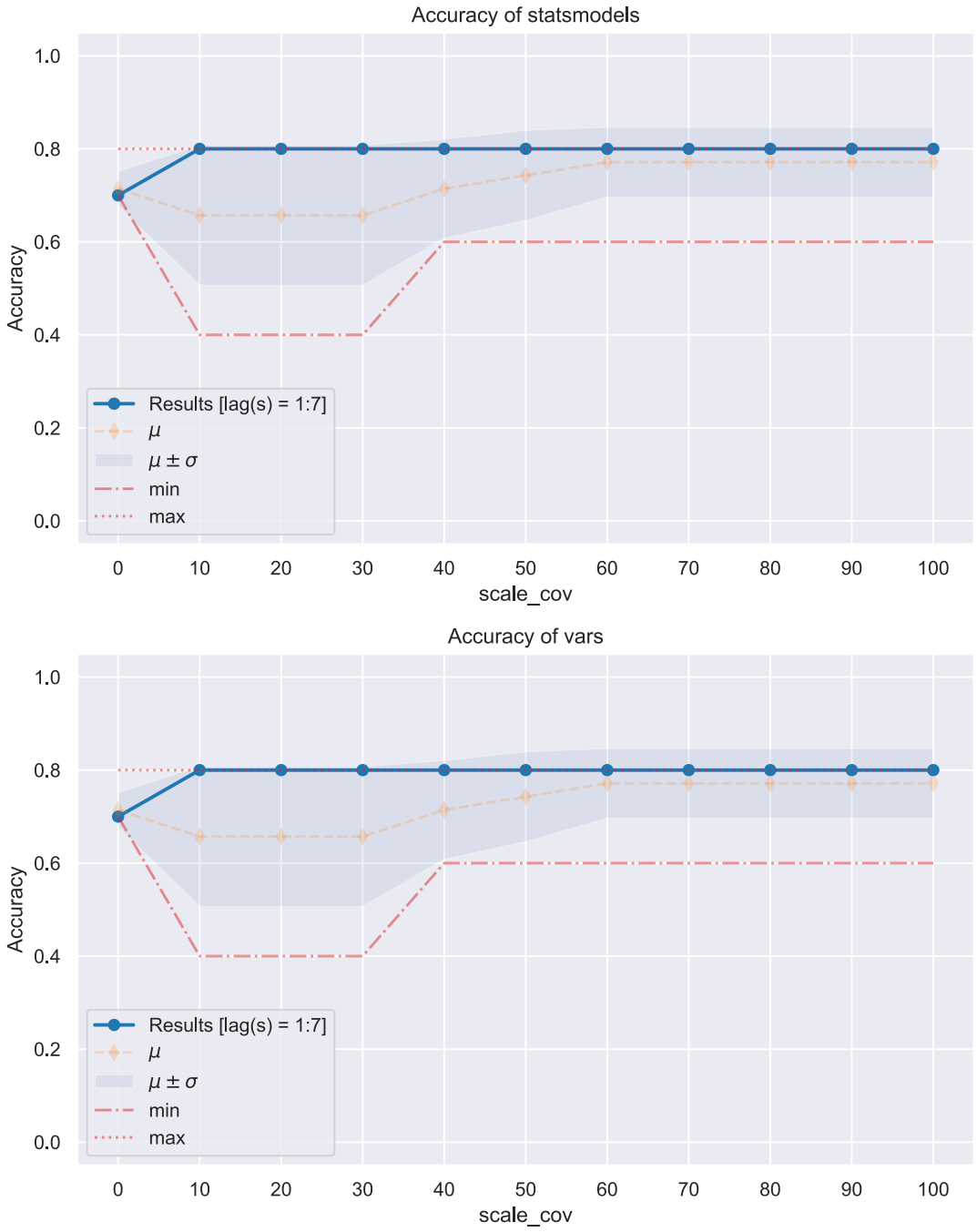


Figure D.8: Accuracy plots (all on one | sensitivity analysis | covariance upscaling)

D.3 Tables

D.3.1 Single analysis

D.3.1.1 Init

	Mean	Std.	Min	Max
`NlinTS`	0.10	0.13	−0.15	0.19
`VLTimeCausality`	−0.56	0.13	−0.83	−0.47
`lmtest`	−0.46	0.09	−0.63	−0.37
`statsmodels`	−0.45	0.11	−0.62	−0.32
`vars`	−0.46	0.09	−0.63	−0.37
All packages	−0.37	0.26	−0.83	0.19

Table D.1: Summary statistics of the Pearson correlation (time series with default settings)

D.3.1.2 Correlated time series

	Mean	Std.	Min	Max
`NlinTS`	−0.01	0.03	−0.07	0.00
`VLTimeCausality`	−0.58	0.15	−0.91	−0.48
`lmtest`	−0.38	0.02	−0.40	−0.35
`statsmodels`	−0.31	0.13	−0.52	−0.13
`vars`	−0.38	0.02	−0.40	−0.35
All packages	−0.33	0.21	−0.91	0.00

Table D.2: Summary statistics of the Pearson correlation (correlated time series)

D.3.1.3 Time series with structural breaks

	Mean	Std.	Min	Max
`NlinTS`	0.01	0.03	−0.01	0.08
`VLTimeCausality`	−0.42	0.12	−0.68	−0.35
`lmtest`	−0.13	0.07	−0.18	0.03
`statsmodels`	−0.32	0.03	−0.35	−0.27
`vars`	−0.10	0.08	−0.18	0.03
All packages	−0.19	0.17	−0.68	0.08

Table D.3: Summary statistics of the Pearson correlation (time series with structural breaks)

D.3.1.4 Time series with a time trend

	Mean	Std.	Min	Max
`NlinTS`	0.00	0.00	0.00	0.00
`VLTimeCausality`	−0.03	0.07	−0.17	0.03
`lmtest`	−0.03	0.07	−0.17	0.03
`statsmodels`	−0.20	0.09	−0.39	−0.13
`vars`	−0.03	0.07	−0.17	0.03
All packages	−0.06	0.10	−0.39	0.03

Table D.4: Summary statistics of the Pearson correlation (time series with a time trend)

Bibliography

Amornbunchornvej, C., Zheleva, E., & Berger-Wolf, T. (2021). Variable-lag granger causality for time series analysis. *ACM Transactions on Knowledge Discovery from Data*, *15*(4(67)), 1–30.

Bao, H.-W.-S. (2022). Brucer: Broadly useful convenient and efficient r functions.

Florens, J. P., & Mouchart, M. (1982). A note on noncausality. *Econometrica*, *50*(3), 583–591.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.

Granger, C. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329–352.

Granger, C. (1988). Some recent development in a concept of causality. *Journal of Econometrics*, *39*(1), 199–211.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics*. McGraw-Hill/Irwin.

Guo, S., Seth, A. K., Kendrick, K. M., Zhou, C., & Feng, J. (2008). Partial granger causality—eliminating exogenous inputs and latent variables. *Journal of Neuroscience Methods*, *172*(1), 79–93.

Hmamouche, Y. (2020). NlinTS: An R Package For Causality Detection in Time Series. *The R Journal*, *12*(1), 21–31.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

Lütkepohl, H., & Reimers, H.-E. (1992). Granger-causality in cointegrated var processes the case of the term structure. *Economics Letters*, *40*(3), 263–268.

Pfaff, B. (2008a). *Analysis of integrated and cointegrated time series with r* (Second) [ISBN 0-387-27960-1]. Springer.

Pfaff, B. (2008b). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4).

Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.*, *85*, 461–464.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

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

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

StataCorp. (2021). *Time-series reference manual*. Stata Press.

Stock, J. H., & Watson, M. W. (2020). *Introduction to econometrics*. Pearson Education Limited.

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, *66*(1), 225–250.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.

Statutory declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Passau, February 11, 2022

(Florian Peschke)