A new prediction model for GDP using Granger lag causality and partial correlation

Vasiliki Sarantidou[1], Christina Karamichalakou[2], Dimitris Kugiumtzis[3]

[1] *Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece*

[2] *Business Statistics Division - Hellenic Statistical Authority, Piraeus, Greece*

[3] *Department of Electrical and Computer Engineering, University of Thessaloniki, Thessaloniki, Greece*

Abstract

The Gross Domestic Product (GDP) is one of the most well-established, well-known and relevant official statistical metrics. The exploration of the components that mainly affect the evolvement of GDP, aiming at concluding to the prediction of GDP, is placed high at the statistical and macroeconomic scientific agenda. Given quarterly time series measurements of GDP and its components, the objective is to predict GDP given the past data up to the current quarter. For many components and long horizon of past values, this is a high-dimensional regression problem, and dimension reduction has to be called in. There are different approaches in the literature for variable selection, such as the least absolute shrinkage and selection operator (LASSO), or variable extraction, such as the principal component regression (PCR). Here, a new prediction model is proposed applying a stepwise forward selection algorithm using as selection criterion the partial correlation for evaluating the conditional lag Granger causality of any of the candidate components (including past GDP) to the next quarter GDP. The termination criterion is a properly designed parametric hypothesis test, ensuring a balance between model complexity and predictive power. A simulation study is conducted to assess the reliability and consistency of the algorithm and compare it to other approaches, such as LASSO and PCR. The applicability of the proposed algorithm is demonstrated using time series data of the Greek GDP and its components, compiled by ELSTAT. By applying the algorithm to this dataset, the variables that most influence GDP fluctuations are identified. The selected variables are then used to form a prediction model, contributing to accurate predictions. This study is carried out in the framework of the EMOS programme of the Aristotle University of Thessaloniki in Greece.

Keywords: Gross Domestic Product, prediction model, multivariate time series, variable selection, partial correlation

1. Introduction

Most economic systems describe interactions among various economic indicators. Multivariate time series analysis using vector autoregressive (VAR) models investigates these interactions for accurate predictions. However, when the multivariate time series has high dimension (many observed variables), dimension reduction techniques are required such as variable extraction and variable selection (Siggiridou and Kugiumtzis, 2016; Dallakyan et al., 2022). Variable selection aims to identify the optimal subset of the observed variables to predict the response variable. Here, we propose a new variable selection algorithm for linear models making use of standard statistical tools, i.e., the partial correlation for measuring interaction and parametric hypothesis test for its significance. This method is applied to Greek GDP data and its components data, compiled by the Hellenic Statistical Authority (ELSTAT) and available on their official website. We focus on the components of the production and expenditure approaches, totaling 20 components. A detailed description is provided in Appendix A. Using a VAR(10) model on 21 variables, including lag values of GDP, we check relationships across 10 lags with final objective to predict GDP.

The paper is structured as follows. In Section 2 the theoretical framework is briefly given. In Section 3 the proposed method is presented, and in Section 4 alternative models of dimension reduction are discussed. The simulation study and the application to Greek GDP data are described and the results are discussed in Section 5. The paper concludes in Section 6 with final remarks.

2. Theoretical Framework

Correlation analysis quantifies relationships between variables with *Pearson correlation coefficient* (denoted r) capturing linear correlations. For continuous variables X and Y and a bivariate sample of size N it is defined as $r(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)\mathrm{var}(Y)}$, were $cov(X,Y)$ denotes the sample covariance and var(*X*) the sample variance of *X*. A parametric significance test for r is designed using the Fisher's transform so that the test statistic r_f has asymptotically normal null distribution with mean 0 and variance $\frac{1}{N-3}$ (Choi et al., 2020). For time series observations of the two variables, $\{X_t, Y_t\}, t = 1, .., N$, the $r(X_t, Y_{t+\tau})$ is the cross-correlation at lag τ.

The *partial correlation coefficient* evaluates the linear relationship between two variables X_1 and X_2, while controlling for other variables $Z = \{Z_1, \dots, Z_m\}$. It is calculated by regressing X_1 and X_2 on Z obtaining residuals u_{X_1} and u_{X_2}. The partial correlation is then $r(XY|Z)_{\square} = corr(u_{X_1}, u_{X_2}) = \frac{\mathrm{cov}(u_{X_1}, u_{X_2})}{\sqrt{\mathrm{var}(u_{X_1})} \cdot \sqrt{\mathrm{var}(u_{X_2})}}$ (Li et al., 2017). Fisher's transformation is also used here to form normal null distribution with mean 0 and variance $\frac{1}{N-m-3}$, where *m* is the number of variables in vector Z (Choi et al., 2020; Williams & Rast, 2020).

3. Proposed method

Let $\{X_{1,t}, \dots, X_{K,t}\}, t = 1, .., N$, be a K-dimensional multivariate time series of length N. The VAR model of order p for the response variable $X_{i,t}$ represents each variable as a linear function of past values of all K variables up to lag p and a white noise error term (Lütkepohl 2005). For large K or p, the VAR(p) model may contain redundant or irrelevant terms and if N is short their coefficients may erroneously be estimated as significant. To render a sparse VAR (error terms much less than Kp), a stepwise variable selection scheme based on the partial correlation coefficient is proposed here, aiming to enhance the prediction accuracy.

Initially, we create an extended dataset of Kp lag variables up to order p from the original K variables: $W = \{X_{1,t-1}, \dots, X_{K,t-1}, X_{1,t-2}, \dots X_{K,t-p}\} = \{w_1, \dots, w_{K\cdot p}\}$. Suppose the response variable is $y \coloneqq X_{1,t}$ (or any of the K variables). Starting with an empty vector $\boldsymbol{w}$, lag variables of W are progressively added based on their correlation with y. The correlation is quantified with the correlation coefficient (cross-correlation) $r(y,w)$ in the first step and partial correlation $r(y,w|\boldsymbol{w})$ for the subsequent steps, where w is any of the Kp candidate lag variables. The lag variable maximizing the absolute value of the correlation is selected to be added to $\boldsymbol{w}$ at each step. For the termination criterion checked at each completed step, a proper parametric hypothesis test is designed. To find the distribution of the maximum absolute correlation, we first assume the normal distribution of the Fisher-transformed (partial) correlation coefficients r_f. Denoting M the maximum of $n \le Kp$ absolute (partial) correlation coefficients and assuming them independent, , its distribution is given by (Choi et al., 2020; Coles, 2001):

$$P[M \le z] = P[r_1 \le z, \dots, r_n \le z] = P[r_1 \le z] \cdot \dots \cdot P[r_n \le z] = \left\{\Phi\left(\frac{z}{s}\right)\right\}^n.$$

Therefore, the critical value z_{1-a} for a significance level a is $z_{1-a} = \Phi^{-1}\left(\sqrt[n]{1-a}\right) \cdot s$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution and s is the standard deviation of the transformed correlation coefficients. At each but the first two steps, a backward revision check is applied (if any of the existing terms in $\boldsymbol{w}$ has to be dropped in view of the selected lag variable) to ensure the selected variables significantly affect the response variable (Derksen & Keselman, 1992).

A pseudo-code of the algorithm is given in Figure 5 in Appendix B.

The steps of the algorithm are illustrated below for the following VAR(2) stationary system on three variables:

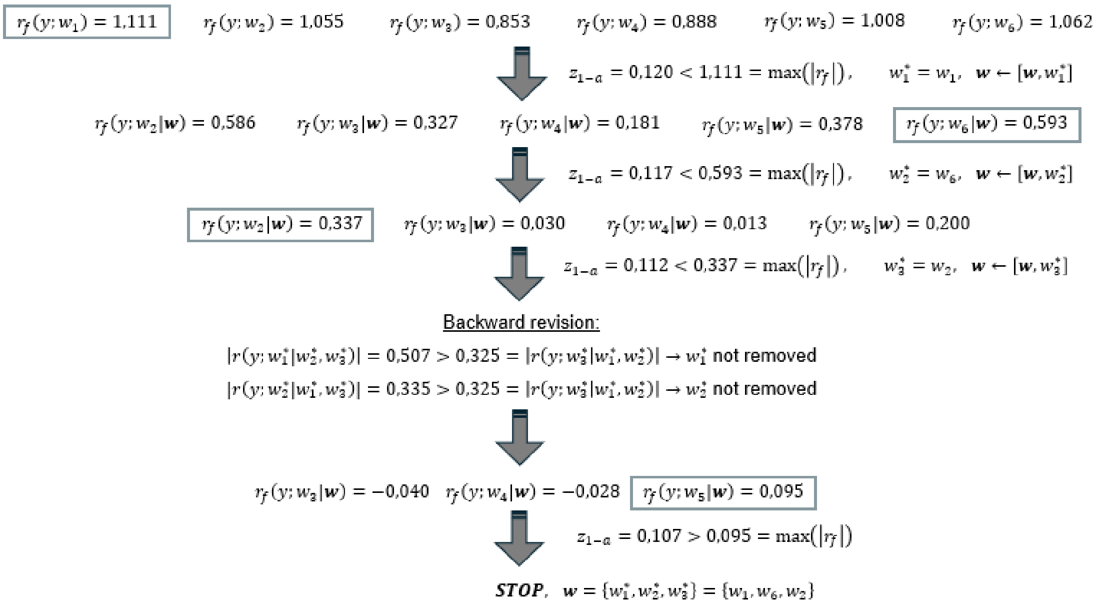
$$X_{1,t} = 0{,}439 X_{1,t-1} + 0{,}294 X_{2,t-1} + 0{,}329 X_{3,t-2} + \varepsilon_{1,t}$$

$$X_{2,t} = 0{,}293 X_{1,t-2} + 0{,}377 X_{2,t-1} + 0{,}316 X_{3,t-1} + \varepsilon_{2,t}$$

$$X_{3,t} = 0{,}191 X_{1,t-1} + 0{,}485 X_{2,t-1} + 0{,}362 X_{3,t-1} - 0{,}179 X_{3,t-2} + \varepsilon_{3,t},$$

where $\varepsilon_{i,t}$, $i = 1,2,3$, is uncorrelated white noise. The response variable is $y \coloneqq X_{1,t}$, and the generated time series has length $N = 400$. In the first step, the cross-correlation between y and each candidate lag variable from the set $W = \{X_{1,t-1}, X_{2,t-1}, \dots, X_{3,t-2}\} = \{w_1, \dots, w_6\}$ is computed and after applying Fisher's transformation (r_f), the lag variable with the highest absolute correlation is selected and tested for significance (coefficient is compared to the respective critical value z_{1-a} for $a = 0.05$). This selected variable is added to the initially empty $\boldsymbol{w}$. Partial correlations are then calculated sequentially, following the same steps until no further lag variables of statistically significant maximum partial correlation are found. Backward revision at each step > 2 ensures optimal variable selection. The procedure is called PartialCor and it is shown analytically in Figure 1 and shows that the algorithm has correctly selected the variables that most affect the response variable: $\boldsymbol{w} = \{w_1, w_6, w_2\} = \{X_{1,t-1}, X_{3,t-2}, X_{2,t-1}\}$.

Figure 1: Steps of the proposed algorithm applied to a VAR(2) system.


$r_f(y; w_1) = 1{,}111$
$r_f(y; w_2) = 1{,}055$
$r_f(y; w_3) = 0{,}853$
$r_f(y; w_4) = 0{,}888$
$r_f(y; w_5) = 1{,}008$
$r_f(y; w_6) = 1{,}062$
$z_{1-a} = 0{,}120 < 1{,}111 = \max(|r_f|)$, $w_1^* = w_1$, $\boldsymbol{w} \leftarrow [\boldsymbol{w}, w_1^*]$
$r_f(y; w_2|\boldsymbol{w}) = 0{,}586$
$r_f(y; w_3|\boldsymbol{w}) = 0{,}327$
$r_f(y; w_4|\boldsymbol{w}) = 0{,}181$
$r_f(y; w_5|\boldsymbol{w}) = 0{,}378$
$r_f(y; w_6|\boldsymbol{w}) = 0{,}593$
$z_{1-a} = 0{,}117 < 0{,}593 = \max(|r_f|)$, $w_2^* = w_6$, $\boldsymbol{w} \leftarrow [\boldsymbol{w}, w_2^*]$
$r_f(y; w_2|\boldsymbol{w}) = 0{,}337$
$r_f(y; w_3|\boldsymbol{w}) = 0{,}030$
$r_f(y; w_4|\boldsymbol{w}) = 0{,}013$
$r_f(y; w_5|\boldsymbol{w}) = 0{,}200$
$z_{1-a} = 0{,}112 < 0{,}337 = \max(|r_f|)$, $w_3^* = w_2$, $\boldsymbol{w} \leftarrow [\boldsymbol{w}, w_3^*]$
Backward revision:
$|r(y; w_1^*|w_2^*, w_3^*)| = 0{,}507 > 0{,}325 = |r(y; w_3^*|w_1^*, w_2^*)| \rightarrow w_1^*$ not removed
$|r(y; w_2^*|w_1^*, w_3^*)| = 0{,}335 > 0{,}325 = |r(y; w_3^*|w_1^*, w_2^*)| \rightarrow w_2^*$ not removed
$r_f(y; w_3|\boldsymbol{w}) = -0{,}040$
$r_f(y; w_4|\boldsymbol{w}) = -0{,}028$
$r_f(y; w_5|\boldsymbol{w}) = 0{,}095$
$z_{1-a} = 0{,}107 > 0{,}095 = \max(|r_f|)$
$\boldsymbol{STOP}$, $\boldsymbol{w} = \{w_1^*, w_2^*, w_3^*\} = \{w_1, w_6, w_2\}$

4. Other methods

This new method is compared with other dimension reduction techniques: 1) LASSO selects variables by shrinking coefficients to zero through a penalty term, with the lambda parameter chosen via five-fold cross-validation (Tibshirani, 1996). 2) The modified Backward in Time Selection (mBTS) builds a dynamic regression model by selecting the most relevant lagged variables to the response starting from the most current lag and going backwards in time (Siggiridou & Kugiumtzis, 2016). 3) PCR is a variable extraction technique that forms the K principal components to be linear combinations of the K original variables ranked under uncorrelatedness and maximum variance assumption, with the number of the first principal components chosen to explain 95% variance of the original data (Massy, 1965). 4) Additionally, the simple autoregressive AR(p) model is applied (Wei, 1990).

5. Results

In this section, we present results for both the simulation study and the application to real data.

5.1 Simulation study

5.1.1 Statistical Evaluation and Setup

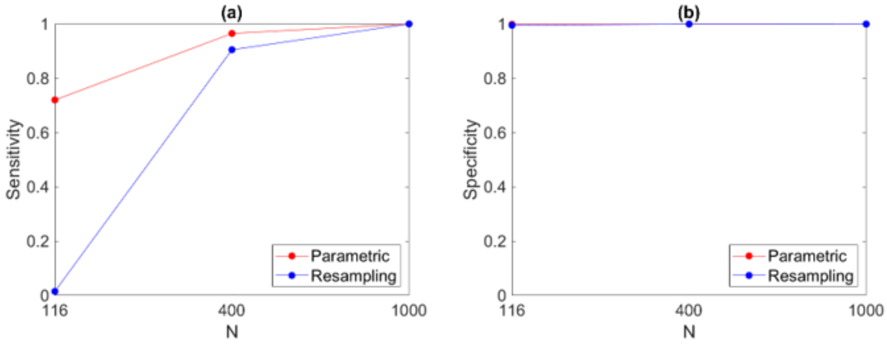
The efficacy of the proposed method is evaluated through simulation assessing the following five metrics. *Sensitivity* measures the proportion of correctly identified relevant variables, *specificity* measures the correctly excluded irrelevant variables and *precision* measures the accuracy of the algorithm’s choices. *Matthews Correlation Coefficient* (MCC) ranging from -1 to +1, indicates binary classification performance, while *F-score* is the harmonic mean of precision and sensitivity (Chicco & Jurman, 2020). The predictive capability is tested with a 70-30% split of the synthetic data into training and test sets and evaluated by the adjusted R^2, denoted R^2_{adj}. Additionally, the robustness of the method’s parametric hypothesis test is compared with a resampling significance test where the resampled time series are derived by random time-shifting.

The simulation systems VAR stationary stochastic processes on 21 variables of order 10, formed on the basis of the Greek GDP time series as follows. First, the proposed algorithm is applied to the dataset three times at a varying significance level a, to attain VAR models of varying sparsity. Each of the three VAR models is then used as the generating stochastic process and 100 realizations are generated. All the subsequent hypothesis tests maintain a significance level of $a = 0.05$.

5.1.2 System 1: Very Sparse

Firstly, to derive a very sparse VAR process significance level is set to $a = 0.05$ (the system equations are given in Appendix C.1). We assess the termination criteria in PartialCor with the parametric test versus the resampling test. For long time series, both criteria perform similarly well in selecting the true lag variables, as shown in Figure 2a and\ excluding irrelevant variables correctly, as shown in Figure 2b (the results on all metrics are given in Appendix C.1).

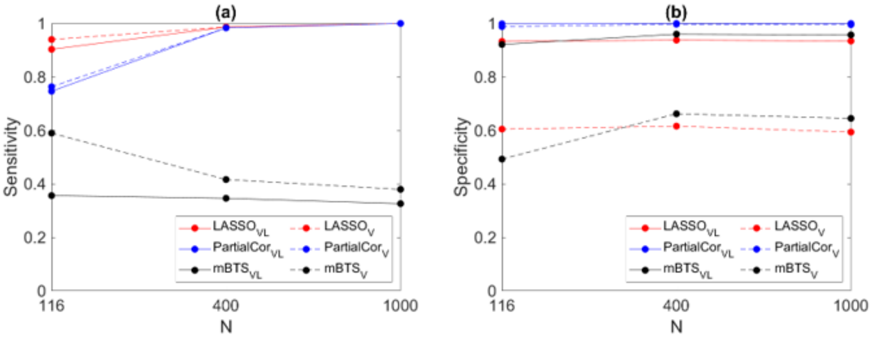
Figure 2: Sensitivity in (a) and Specificity in (b) vs time series length N for the termination criterion using the parametric and resampling significance test as given in the legend.



5.1.3 System 2: Sparse

Increasing the significance level to $a = 0.25$ in our procedure, the generating stochastic process contain more terms (system equations are given in Appendix C.2). We compare the proposed algorithm with LASSO and mBTS for variable selection, in two stages: ’*VL*’, evaluating the ability to select both the correct variables and their appropriate lags, and ‘*V*’, assessing the selection of the appropriate variables alone regardless of the lag. Figure 3a shows high sensitivity for PartialCor and LASSO and Figure3b shows high specificity for PartialCor for both stages, whereas LASSO and mBTS perform poorly when assessing stage ‘V’ selection. Table 4 in Appendix C.2 presents results for all five metrics.

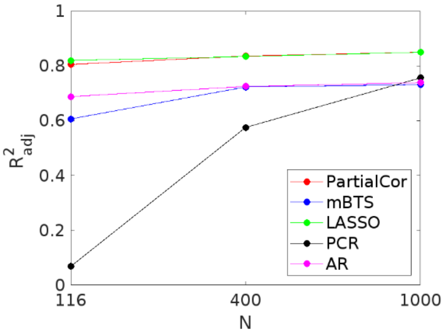
Figure 3: Sensitivity in (a) and Specificity in (b) vs N for the variable selection methods PartialCor, LASSO and mBTS in the two stages VL and L as denoted in the legend.



5.1.4 System 3: Dense

For significance level $a = 0.40$ we obtain the largest VAR system (system equations are given in Appendix C.2). We compare the performance of PartialCor in terms of R^2_{adj} with this of LASSO, mBTS, PCR and AR(10). As shown in Figure 4, regardless of N, LASSO and PartialCor exhibit nearly identical predictive power. Detailed results on the predictive performance for each of the 21 response variables are given in Table 5 in Appendix C.3.

Figure 4: Average R^2_{adj} vs time series length N over 100 realizations for five methods listed in the legend considering $X_{1,t}$ as the response variable



5.2 Greek GDP data

We aim to predict quarterly GDP for the next quarter using historical data, including GDP past values from a sparse VAR(10) model. The proposed algorithm PatialCor identified two significant lag variables, 'GDP_{t-1}' and 'subsidies_{t-6}'. Compared to LASSO and mBTS, our algorithm selects fewer optimal lag variables. LASSO method includes these two lag variables and 27 more, while mBTS includes 'GDP_{t-1}', a different lag for the other and 25 more, only 5 of these common with LASSO (the exact subsets are given in Table 6 in Appendix D.1). We then performed linear regression with the selected lag variables making both in-sample and five-fold cross-validation prediction. For this, we use R^2_{adj} and also the normalized root mean square error (NRMSE). LASSO performs best in in-sample predictions as it achieves the smallest NRMSE, while PartialCor scores better in NRMSE than PCR and AR(10) methods. Cross-validation results show that PartialCor outperforms the other four methods giving the highest R^2_{adj} and smallest NRMSE being strongly statistically significant using the Diebold Mariano test (Chen et al., 2014). Actually this test found no statistical difference of PartialCor only to AR(10) in in-sample prediction, as shown in Table 1. Time comparison shows similar performance between the variable selection methods, with PartialCor being slightly faster.

Table 1: NRMSE, R^2_{adj} and execution time in seconds for GDP prediction from 1995 Q1 to 2023 Q4. The (*) denotes no statistical difference in NRMSE from PartialCor using Diebold Mariano test.

Prediction	Metric	PartialCor	mBTS	LASSO	PCR	AR(10)
In-Sample	NRMSE	0,143	0,114	**0,098**	0,228	0,152*
	R^2_{adj}	0,979	0,982	**0,986**	0,945	0,974
	Time	3,803	5,677	5,079	0,007	0,001
Cross-Validation	NRMSE	**0,192**	0,820	0,276	0,626	0,254
	R^2_{adj}	**0,962**	0,314	0,923	0,601	0,934
	Time	3,811	5,687	5,088	0,053	0,008

The reference year for GDP and its components was updated to 2015, impacting data from 2010 to 2023. While the initial dataset starts from 1995, revisions from 2010 onwards reflect the new reference year of 2015. Therefore, we repeat the predictions limiting the dataset to this period. The PartialCor identifies '$\mathrm{GDP}_{\mathrm{t-1}}$' as significant variable, similar to LASSO and mBTS. While LASSO selects 14 variables, mBTS finds 46, neither of them common with these of LASSO (the subsets of lag variables are given in Table 7 in Appendix D.2).

Table 2: NRMSE, R^2_{adj} and execution time in seconds for GDP prediction from 2010 Q1 to 2023 Q4 (revised time series). The (*) denotes no statistical difference in NRMSE from PartialCor using Diebold Mariano test.

Prediction	Metric	PartialCor	mBTS	LASSO	PCR	AR(10)
In-Sample	NRMSE	0,559	**0,000**	0,246*	0,459*	0,548*
	R^2_{adj}	0,665	**1,000**	0,907	0,706	0,616
	Time	6,484	7,445	6,955	0,209	0,006
Cross-Validation	NRMSE	0,674	5,372	**0,666***	0,767*	0,838*
	R^2_{adj}	0,526	-29,168	**0,537**	0,385	0,266
	Time	6,688	8,026	6,286	0,019	0,008

In Table 2, in-sample predictions show mBTS are excellent but due to overfitting (many lag variables) as it fails in cross-validation. Other methods demonstrate decreased performance due to fewer observations. Regarding PartialCor, even though it obtains the worst results in the in-sample predictions, the Diebold-Mariano test showed that it has no statistical difference from all other methods except from mBTS. In cross-validation there is also no statistical difference with all models apart from mBTS.

6. Conclusions

In this work we introduced a new method for GDP prediction using Ganger causality and partial correlation-based lag variable selection called PartialCor. This approach enhances multivariate time series management and forecasting by finding direct linear relationships between lag variables and response. When the PartialCor was applied to Greek GDP data and its components it identified two variables as optimal, ‘GDP’ for the previous quarter and ‘subsidies’ of 6 quarters before. Restricting data to revised quarters highlighted ‘GDP_{t-1}’ as the sole significant variable. Comparative analysis, including simulations and real-word application to Greek GDP data, highlights that the proposed algorithm PartialCor is a simple, fast and accurate method for sparse modeling and prediction, positioning it as a compelling choice among other more involved methods. Explicitly, this method includes a reliable parametric hypothesis-based termination criterion and offers straightforward interpretation, effective and consistent performance in variable selection as well as it is time efficient.

References

Chen, H., Wan, Q., & Wang, Y. (2014). Refined diebold-mariano test methods for the evaluation of wind power forecasting models. *Energies*, *7*(7), 4185–4198. https://doi.org/10.3390/en7074185

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Choi, D., Li, L., Liu, H., & Zeng, L. (2020). A recursive partitioning approach for subgroup identification in brain–behaviour correlation analysis. *Pattern Analysis and Applications*, *23*(1), 161–177. https://doi.org/10.1007/s10044-018-00775-y

Coles, S., Bawa, J., Trenner, L., Dorazio, P. (2001). An introduction to statistical modeling of extreme values. Volume 208. Springer.

Dallakyan, A., Kim, R., Pourahmadi, M. (2022) Time series graphical lasso and sparse VAR estimation, *Computational Statistics & Data Analysis*, 176, 107557. https://doi.org/10.1016/j.csda.2022.107557

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282. https://doi.org/10.1111/j.2044-8317.1992.tb00992.x

Li, R., Liu, J., & Lou, L. (2017). Variable selection via partial correlation. *Statistica Sinica*, *27*(3), 983–996. https://doi.org/10.5705/ss.202015.0473

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

Massy, W. F. (1965). Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*, *60*(309), 234–256. https://doi.org/10.2307/2283149

Siggiridou, E., & Kugiumtzis, D. (2016). Granger Causality in Multivariate Time Series Using a Time-Ordered Restricted Vector Autoregressive Model. *IEEE Transactions on Signal Processing*, *64*(7), 1759–1773. https://doi.org/10.1109/TSP.2015.2500893

The MathWorks Inc. (2023). MATLAB version: 23.2.0 (R2022b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wei, W. W. S. (1990). Time series analysis: univariate and multivariate methods. Pearson Addison Wesley, Redwood City.

Williams, D. R., & Rast, P. (2020). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 187–212. https://doi.org/10.1111/bmsp.12173

Appendix

Appendix A: Variables Description

National Accounts, compiled by Hellenic Statistical Authority (ELSTAT) following the European System of Accounts ESA 2010, provide a comprehensive description of a country's economy Gross Domestic Product (GDP) being a key indicator, represents total value of final goods and services produced, without intermediate consumption. The estimations we use are chain link volume data, seasonally adjusted based on JDemetra+ software in accordance with the Greek diary. Measurements recorded start from the first quarter of 1995 up to the most recent quarter. Listed below are the components of production and expenditure approaches that will be used to predict GDP.

From production approach mainly contains classes of economic activities, which European Community (NACE Rev.2) in ten cumulative sections, there are the following components.

- *sectionA*. It includes agricultural, forestry and fishing occupations.
- *sectionBCDE*. This one contains the extraction of natural minerals as well as activities aimed at preparing the raw materials so that they can be sold on the market. Additionally, it includes manufacturing and the operation of business responsible for the supply of electricity, natural gas, steam and air conditioning. Finally, activities related to collection, treatment and supply of water and treatment process of all types of waste are also included in this component.
- *sectionF*. It encompasses construction activities such as, construction of buildings or construction of civil engineering projects and their repair.
- *sectionGHI*. This component includes retail and wholesale trade. It also includes the trade, repair and washing of vehicles and motorcycles, activities of transporting goods and people through all kinds of roads. Finally, it contains the provision of short-term accommodation services and catering services for immediate consumption.
- *sectionJ*. This section encompasses various media, information and communication technologies as well as all the activities of production and distribution of information products.
- *sectionK*. It includes financial services and insurance activities.
- *sectionL*. Activities related to the management of real estate belong to this component.
- *sectionMN*. It concerns activities of professional, scientific and technical content as well as activities that support various business pursuits.
- *sectionOPQ*. This component contains government-related and compulsory social security activities. Also, all education stages, activities related to human health protection issues, provided by professionals, and social welfare issues are included.
- *sectionRSTU*. Related to entertainment and arts. Additionally, it contains activities related to the repair of household goods and services that are not classified in any other category.
- *taxes*. Taxes on products.
- *subsidies*. Subsidies on products.

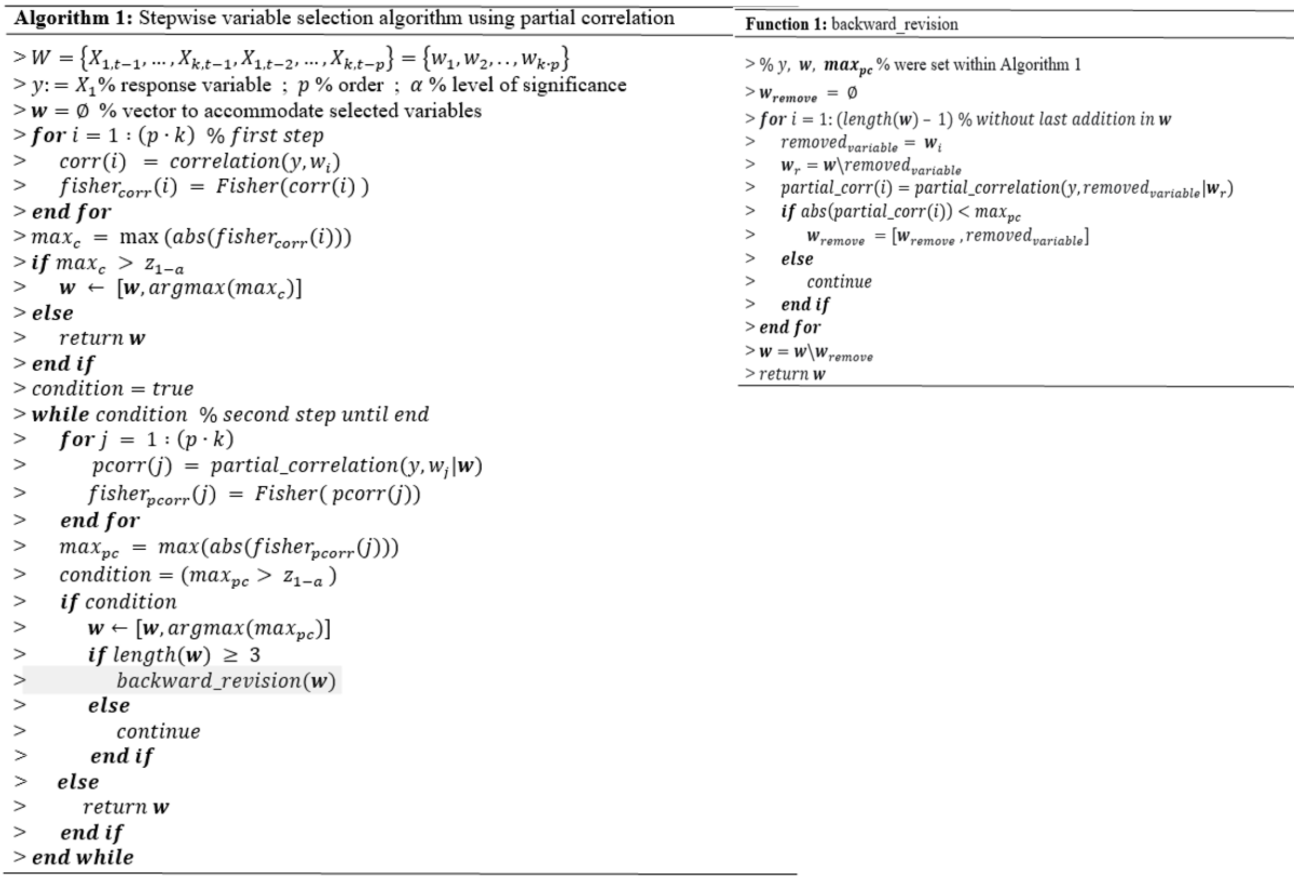
From expenditure approach, there are the following components.

- *households_NPISH*. It refers to expenditure of households and NPISH (Non-profit institutions that serve households), which are neither financed nor controlled by the state.
- *GG*. It contains General Government expenditure.
- *fixed_gross_capital_formation*. Gross fixed capital formation is the country’s investments in fixed assets (i.e. physical assets with a lifespan of more than one year).
- *gross_capital_formation*. It is the sum of gross fixed capital formation that was defined in the previous component with the value of goods held as inventory by businesses.
- *export_goods*. Exports of goods.
- *export_services*. Exports of services.
- *import_goods*. Imports of goods.
- *import_services*. Imports of services.

Thus, 20 components are linked to GDP. More information can be found on the website of ELSTAT (https://www.statistics.gr/) and the European Statistical Service (https://ec.europa.eu/eurostat).

Appendix B: Proposed Algorithm

Figure 5: Pseudo-code including the steps of the proposed algorithm PartialCor to find the optimal subset of lag variables for prediction.



Algorithm 1: Stepwise variable selection algorithm using partial correlation

```
> $W = \{X_{1,t-1}, \dots, X_{k,t-1}, X_{1,t-2}, \dots, X_{k,t-p}\} = \{w_1, w_2, \dots, w_{k \cdot p}\}$
> $y := X_1$ % response variable ; $p$ % order ; $\alpha$ % level of significance
> $\mathbf{w} = \emptyset$ % vector to accommodate selected variables
> for $i = 1 : (p \cdot k)$ % first step
>     $corr(i) = correlation(y, w_i)$
>     $fisher_{corr}(i) = Fisher(corr(i))$
> end for
> $max_c = \max(abs(fisher_{corr}(i)))$
> if $max_c > z_{1-a}$
>     $\mathbf{w} \leftarrow [\mathbf{w}, argmax(max_c)]$
> else
>     return $\mathbf{w}$
> end if
> $condition = true$
> while condition % second step until end
>     for $j = 1 : (p \cdot k)$
>         $pcorr(j) = partial\_correlation(y, w_j | \mathbf{w})$
>         $fisher_{pcorr}(j) = Fisher(pcorr(j))$
>     end for
>     $max_{pc} = max(abs(fisher_{pcorr}(j)))$
>     $condition = (max_{pc} > z_{1-a})$
>     if condition
>         $\mathbf{w} \leftarrow [\mathbf{w}, argmax(max_{pc})]$
>         if $length(\mathbf{w}) \geq 3$
>             $backward\_revision(\mathbf{w})$
>         else
>             continue
>         end if
>     else
>         return $\mathbf{w}$
>     end if
> end while
```

Function 1: backward_revision

```
> % $y$, $\mathbf{w}$, $\mathbf{max_{pc}}$ % were set within Algorithm 1
> $\mathbf{w}_{remove} = \emptyset$
> for $i = 1 : (length(\mathbf{w}) - 1)$ % without last addition in $\mathbf{w}$
>     $removed_{variable} = \mathbf{w}_i$
>     $\mathbf{w}_r = \mathbf{w} \backslash removed_{variable}$
>     $partial\_corr(i) = partial\_correlation(y, removed_{variable} | \mathbf{w}_r)$
>     if $abs(partial\_corr(i)) < max_{pc}$
>         $\mathbf{w}_{remove} = [\mathbf{w}_{remove}, removed_{variable}]$
>     else
>         continue
>     end if
> end for
> $\mathbf{w} = \mathbf{w} \backslash \mathbf{w}_{remove}$
> return $\mathbf{w}$
```

Appendix C: Simulation Study Systems and Results

Appendix C.1: System 1: Very Sparse

A VAR(10) stationary process on 21 variables derived by applying the proposed algorithm PartialCor in Greek GDP data while setting the significance level $\alpha = 0.05$.

$$X_{1,t} = a_{1,1}X_{1,t-1} + a_{1,2}X_{13,t-6} + e_{1,t}$$

$$X_{2,t} = a_{2,1}X_{2,t-1} + e_{2,t}$$

$$X_{3,t} = a_{3,1}X_{3,t-1} + e_{3,t}$$

$$X_{4,t} = a_{4,1}X_{4,t-1} + e_{4,t}$$

$$X_{5,t} = a_{5,1}X_{5,t-1} + a_{5,2}X_{4,t-4} + e_{5,t}$$

$$X_{6,t} = a_{6,1}X_{6,t-1} + a_{6,2}X_{12,t-4} + a_{6,3}X_{2,t-10} + e_{6,t}$$

$$X_{7,t} = a_{7,1}X_{7,t-1} + e_{7,t}$$

$$X_{8,t} = a_{8,1}X_{2,t-1} + a_{8,2}\, X_{8,t-1} + a_{8,3}X_{8,t-2} + e_{8,t}$$

$$X_{9,t} = a_{9,1}X_{9,t-1} + a_{9,2}X_{13,t-6} + e_{9,t}$$

$$X_{10,t} = a_{10,1}X_{10,t-1} + a_{10,2}X_{1,t-4} + a_{10,3}X_{2,t-9} + a_{10,4}X_{4,t-9} + a_{10,5}X_{9,t-9} + e_{10,t}$$

$$X_{11,t} = a_{11,1}X_{11,t-1} + a_{11,2}X_{16,t-2} + a_{11,3}X_{4,t-9} + a_{11,4}X_{17,t-10} + e_{11,t}$$

$$X_{12,t} = a_{12,1}X_{6,t-1} + a_{12,2}X_{12,t-1} + e_{12,t}$$

$$X_{13,t} = \alpha_{13,1}X_{13,t-1} + e_{13,t}$$

$$X_{14,t} = a_{14,1}X_{14,t-1} + a_{14,2}X_{15,t-1} + a_{14,3}X_{5,t-7} + e_{14,t}$$

$$X_{15,t} = a_{15,1}X_{15,t-1} + a_{15,2}X_{16,t-2} + e_{15,t}$$

$$X_{16,t} = a_{16,1}X_{16,t-1} + e_{16,t}$$

$$X_{17,t} = a_{17,1}X_{16,t-1} + a_{17,2}X_{17,t-1} + e_{17,t}$$

$$X_{18,t} = a_{18,1}X_{18,t-1} + a_{18,2}X_{18,t-2} + e_{18,t}$$

$$X_{19,t} = a_{19,1}X_{19,t-1} + e_{19,t}$$

$$X_{20,t} = a_{20,1}X_{20,t-1} + e_{20,t}$$

$$X_{21,t} = a_{21,1}X_{21,t-1} + e_{21,t}$$

Table 3: Mean and standard deviation (std) of five metrics: sensitivity, specificity, precision, MCC and f-score over 100 realizations using parametric test and resampling test (with 100 replications) as termination criterion in the proposed variable selection algorithm for various time series length N, for response variable $X_{1,t}$.

N	Test	Statistic	Sensitivity	Specificity	Precision	MCC	F-Score
116	parametric	mean	0,720	0,999	0,830	0,761	0,751
		std	0,358	0,002	0,345	0,333	0,332
	resampling	mean	0,015	0,995	0,020	0,010	0,017
		std	0,111	0,001	0,141	0,123	0,120
400	parametric	mean	0,965	0,999	0,957	0,957	0,953
		std	0,128	0,002	0,120	0,105	0,112
	resampling	mean	0,905	1,000	0,978	0,934	0,928
		std	0,210	0,001	0,120	0,152	0,161
1000	parametric	mean	1,000	0,999	0,957	0,976	0,973
		std	0,000	0,002	0,120	0,068	0,075
	resampling	mean	1,000	1,000	0,970	0,983	0,982
		std	0,000	0,001	0,096	0,053	0,058

Appendix C.2: System 2: Sparse

As system 1 in Appendix C.1 but for $\alpha = 0.25$.

$X_{1,t} = a_{1,1}X_{1,t-1} + a_{1,2}X_{15,t-1} + a_{1,3}X_{13,t-6} + +e_{1,t}$

$X_{2,t} = a_{2,1}X_{2,t-1} + a_{2,2}X_{19,t-6} + e_{2,t}$

$X_{3,t} = a_{3,1}X_{3,t-1} + e_{3,t}$

$X_{4,t} = a_{4,1}X_{3,t-1} + a_{4,2}X_{4,t-1} + a_{4,3}X_{8,t-10} + e_{4,t}$

$X_{5,t} = a_{5,1}X_{5,t-1} + a_{5,2}X_{4,t-4} + e_{5,t}$

$X_{6,t} = a_{6,1}X_{6,t-1} + a_{6,2}X_{12,t-4} + a_{6,3}X_{2,t-10} + e_{6,t}$

$X_{7,t} = a_{7,1}X_{7,t-1} + a_{7,2}X_{20,t-2} + e_{7,t}$

$X_{8,t} = a_{8,1}X_{2,t-1} + a_{8,2}\ X_{8,t-1} + a_{8,3}X_{19,t-1} + a_{8,4}X_{8,t-2} + e_{8,t}$

$X_{9,t} = a_{9,1}X_{9,t-1} + a_{9,2}X_{12,t-3} + a_{9,3}X_{13,t-6} + e_{9,t}$

$X_{10,t} = a_{10,1}X_{9,t-1} + a_{10,2}X_{10,t-1} + a_{10,3}X_{1,t-4} + a_{10,4}X_{2,t-9} + a_{10,5}X_{4,t-9} + a_{10,6}X_{9,t-9} + e_{10,t}$

$X_{11,t} = a_{11,1}X_{11,t-1} + a_{11,2}X_{16,t-2} + a_{11,3}X_{20,t-3} + a_{11,4}X_{4,t-9} + a_{11,5}X_{10,t-9} + e_{11,t}$

$X_{12,t} = a_{12,1}X_{6,t-1} + a_{12,2}X_{12,t-1} + a_{12,3}X_{3,t-2} + a_{12,4}X_{11,t-3} + a_{12,5}X_{4,t-9} + e_{12,t}$

$X_{13,t} = \alpha_{13,1}X_{13,t-1} + \alpha_{13,2}X_{7,t-2} + \alpha_{13,3}X_{14,t-10} + e_{13,t}$

$X_{14,t} = a_{14,1}X_{3,t-1} + a_{14,2}X_{14,t-1} + a_{14,3}X_{15,t-1} + a_{14,4}X_{5,t-7} + e_{14,t}$

$X_{15,t} = a_{15,1}X_{15,t-1} + a_{15,2}X_{16,t-2} + a_{15,3}X_{17,t-2} + e_{15,t}$

$X_{16,t} = a_{16,1}X_{16,t-1} + a_{16,2}X_{11,t-2} + a_{16,3}X_{2,t-5} + e_{16,t}$

$X_{17,t} = a_{17,1}X_{15,t-1} + a_{17,2}X_{17,t-1} + a_{17,3}X_{16,t-2} + a_{17,4}X_{13,t-7} + a_{17,5}X_{8,t-10} + e_{17,t}$

$X_{18,t} = a_{18,1}X_{18,t-1} + a_{18,2}X_{18,t-2} + e_{18,t}$

$X_{19,t} = a_{19,1}X_{19,t-1} + e_{19,t}$

$X_{20,t} = a_{20,1}X_{20,t-1} + a_{20,2}X_{13,t-9} + e_{20,t}$

$X_{21,t} = a_{21,1}X_{21,t-1} + a_{21,2}X_{18,t-6} + e_{21,t}$

Table 4: Average sensitivity, specificity, precision, MCC and f-score over 100 realizations of system 2 and for the methods Partial Cor, LASSO and mBTS, for various time series lengths *N*, considering $X_{1,t}$ as the response variable.

N	Method	Stage	Sensitivity	Specificity	Precision	MCC	F-Score
116	LASSO	V	0,940	0,605	0,352	0,432	0,482
		VL	0,903	0,933	0,260	0,441	0,369
	PartialCor	V	0,763	0,988	0,908	0,801	0,808
		VL	0,747	0,999	0,875	0,796	0,786
	mBTS	V	0,590	0,493	0,172	0,065	0,256
		VL	0,356	0,922	0,079	0,134	0,123
400	LASSO	V	0,987	0,616	0,346	0,453	0,497
		VL	0,987	0,938	0,244	0,462	0,376
	PartialCor	V	0,983	0,997	0,988	0,982	0,983
		VL	0,983	1,000	0,988	0,984	0,983
	mBTS	V	0,416	0,662	0,185	0,065	0,250
		VL	0,346	0,960	0,134	0,191	0,186
1000	LASSO	V	1,000	0,594	0,348	0,453	0,496
		VL	1,000	0,934	0,243	0,461	0,372
	Partial Cor	V	1,000	0,996	0,981	0,988	0,989
		VL	1,000	1,000	0,976	0,987	0,986
	mBTS	V	0,380	0,645	0,159	0,022	0,220
		VL	0,326	0,957	0,115	0,169	0,164

Appendix C.3: System 3: Dense

As system 1 in Appendix C.1 but for $\alpha = 0.40$..

$X_{1,t} = a_{1,1}X_{1,t-1} + a_{1,2}X_{15,t-1} + a_{1,3}X_{13,t-6} + +e_{1,t}$

$X_{2,t} = a_{2,1}X_{2,t-1} + a_{2,2}X_{19,t-6} + e_{2,t}$

$X_{3,t} = a_{3,1}X_{3,t-1} + e_{3,t}$

$X_{4,t} = a_{4,1}X_{3,t-1} + a_{4,2}X_{4,t-1} + a_{4,3}X_{8,t-10} + e_{4,t}$

$X_{5,t} = a_{5,1}X_{5,t-1} + a_{5,2}X_{4,t-4} + e_{5,t}$

$X_{6,t} = a_{6,1}X_{6,t-1} + a_{6,2}X_{15,t-1} + a_{6,3}X_{12,t-4} + a_{6,4}X_{3,t-8} + a_{6,5}X_{4,t-9} + a_{6,5}X_{15,t-9} + e_{6,t}$

$X_{7,t} = a_{7,1}X_{7,t-1} + a_{7,2}X_{20,t-2} + e_{7,t}$

$X_{8,t} = a_{8,1}X_{2,t-1} + a_{8,2}\ X_{8,t-1} + a_{8,3}X_{19,t-1} + a_{8,4}X_{8,t-2} + e_{8,t}$

$X_{9,t} = a_{9,1}X_{9,t-1} + a_{9,2}X_{2,t-2} + a_{9,3}X_{12,t-3} + a_{9,4}X_{13,t-6} + a_{9,5}X_{4,t-8} + a_{9,6}X_{11,t-8} + e_{9,t}$

$X_{10,t} = a_{10,1}X_{9,t-1} + a_{10,2}X_{10,t-1} + a_{10,3}X_{9,t-2} + a_{10,4}X_{1,t-4} + a_{10,5}X_{4,t-5} + a_{10,6}X_{2,t-9} + a_{10,7}X_{4,t-9} +$
$a_{10,8}X_{9,t-9} + e_{10,t}$

$X_{11,t} = a_{11,1}X_{11,t-1} + a_{11,2}X_{16,t-2} + a_{11,3}X_{20,t-3} + a_{11,4}X_{4,t-9} + a_{11,5}X_{10,t-9} + e_{11,t}$

$X_{12,t} = a_{12,1}X_{6,t-1} + a_{12,2}X_{12,t-1} + a_{12,3}X_{3,t-2} + a_{12,4}X_{11,t-3} + a_{12,5}X_{4,t-9} + e_{12,t}$

$X_{13,t} = \alpha_{13,1}X_{13,t-1} + \alpha_{13,2}X_{7,t-2} + \alpha_{13,3}X_{14,t-10} + e_{13,t}$

$X_{14,t} = a_{14,1}X_{3,t-1} + a_{14,2}X_{14,t-1} + a_{14,3}X_{15,t-1} + a_{14,4}X_{5,t-7} + e_{14,t}$

$X_{15,t} = a_{15,1}X_{15,t-1} + a_{15,2}X_{16,t-2} + a_{15,3}X_{17,t-2} + e_{15,t}$

$X_{16,t} = a_{16,1}X_{16,t-1} + a_{16,2}X_{11,t-2} + a_{16,3}X_{2,t-5} + e_{16,t}$

$X_{17,t} = a_{17,1}X_{15,t-1} + a_{17,2}X_{17,t-1} + a_{17,3}X_{16,t-2} + a_{17,4}X_{13,t-7} + a_{17,5}X_{8,t-10} + e_{17,t}$

$X_{18,t} = a_{18,1}X_{18,t-1} + a_{18,2}X_{18,t-2} + e_{18,t}$

$X_{19,t} = a_{19,1}X_{19,t-1} + a_{19,2}X_{19,t-2} + e_{19,t}$

$X_{20,t} = a_{20,1}X_{20,t-1} + a_{20,2}X_{13,t-9} + e_{20,t}$

$X_{21,t}a_{21,1}X_{21,t-1} + a_{21,2}X_{18,t-6} + e_{21,t}$

Table 5: Average values of R_{adj}^2 from predictions over 100 realizations for time series of length 400 for five models considering each variable of the multivariate time series as the response variable.

Response	PartialCor	mBTS	LASSO	PCR	AR
$X_{1,t}$	**0,836**	0,723	0,834	0,575	0,725
$X_{2,t}$	**0,766**	0,602	**0,766**	0,444	0,542
$X_{3,t}$	0,636	0,574	**0,665**	0,304	0,555
$X_{4,t}$	**0,820**	0,606	0,819	0,554	0,547
$X_{5,t}$	**0,788**	0,602	0,786	0,501	0,510
$X_{6,t}$	**0,915**	0,744	0,911	0,777	0,551
$X_{7,t}$	0,730	0,538	**0,742**	0,422	0,551
$X_{8,t}$	**0,856**	0,571	0,852	0,623	0,499
$X_{9,t}$	**0,912**	0,710	0,907	0,763	0,567
$X_{10,t}$	**0,941**	0,738	0,935	0,828	0,573
$X_{11,t}$	**0,915**	0,666	0,911	0,774	0,542
$X_{12,t}$	**0,914**	0,714	0,912	0,779	0,553
$X_{13,t}$	**0,834**	0,623	0,833	0,591	0,607
$X_{14,t}$	**0,851**	0,724	0,847	0,621	0,595
$X_{15,t}$	**0,878**	0,623	0,876	0,690	0,596
$X_{16,t}$	**0,862**	0,614	0,860	0,661	0,520
$X_{17,t}$	**0,901**	0,672	0,897	0,734	0,574
$X_{18,t}$	0,769	0,496	**0,776**	0,493	0,464
$X_{19,t}$	**0,786**	0,585	**0,786**	0,482	0,503
$X_{20,t}$	0,781	0,617	**0,782**	0,509	0,644
$X_{21,t}$	**0,755**	0,541	**0,755**	0,471	0,530

Appendix D: Greek GDP Data Results

Appendix D.1: Greek GDP Data 1995 Q1 – 2023 Q4

Table 6: Optimal subsets of lag variables found by applying the variable selection methods LASSO and mBTS on Greek GDP data and its components.

LASSO		mBTS	
GDP$_{t-1}$	import_services$_{t-2}$	GDP$_{t-1}$	sectionL$_{t-1}$
sectionBCDE$_{t-1}$	sectionRSTU$_{t-3}$	GDP$_{t-2}$	sectionL$_{t-7}$
sectionF$_{t-1}$	GG$_{t-3}$	GDP$_{t-3}$	sectionMN$_{t-5}$
sectionGHI$_{t-1}$	sectionK$_{t-4}$	GDP$_{t-7}$	sectionOPQ$_{t-2}$
sectionK$_{t-1}$	import_goods$_{t-4}$	sectionA$_{t-2}$	sectionRSTU$_{t-6}$
sectionOPQ$_{t-1}$	sectionJ$_{t-5}$	sectionA$_{t-5}$	Subsidies$_{t-1}$
sectionRST$_{t-1}$	sectionA$_{t-6}$	sectionBCDE$_{t-1}$	Subsidies$_{t-7}$
GG$_{t-1}$	Subsidies$_{t-6}$	sectionBCDE$_{t-2}$	GG$_{t-8}$
import_goods$_{t-1}$	Subsidies$_{t-7}$	sectionBCDE$_{t-3}$	gross_fixed_capital_formation$_{t-3}$
import_services$_{t-1}$	sectionBCDE$_{t-8}$	sectionF$_{t-5}$	gross_fixed_capital_formation$_{t-9}$
sectionA$_{t-2}$	Subsidies$_{t-8}$	sectionGHI$_{t-3}$	export_goods$_{t-2}$
sectionBCDE$_{t-2}$	sectionF$_{t-9}$	sectionGHI$_{t-5}$	export_services$_{t-2}$
sectionOPQ$_{t-2}$	sectionK$_{t-10}$	sectionJ$_{t-7}$	export_services$_{t-10}$
Subsidies$_{t-2}$	sectionRSTU$_{t-10}$	sectionK$_{t-4}$	
import_goods$_{t-2}$			

Appendix D.2: Greek GDP Data 2010 Q1 – 2023 Q4

Table 7: Optimal subsets of lag variables found by applying the variable selection methods LASSO and mBTS on Greek GDP data and its components.

LASSO	**mBTS**	
GDP$_{t-1}$	GDP$_{t-1}$	sectionGHI$_{t-3}$
sectionL$_{t-1}$	GDP$_{t-2}$	sectionGHI$_{t-5}$
gross_capital_formation$_{t-2}$	GDP$_{t-3}$	sectionJ$_{t-6}$
gross_fixed_capital_formation$_{t-2}$	GDP$_{t-6}$	sectionJ$_{t-8}$
import_goods$_{t-2}$	GDP$_{t-7}$	sectionJ$_{t-9}$
GG$_{t-3}$	GDP$_{t-8}$	sectionK$_{t-3}$
gross_capital_formation$_{t-3}$	GDP$_{t-10}$	sectionK$_{t-4}$
gross_fixed_capital_formation$_{t-3}$	sectionA$_{t-1}$	sectionL$_{t-7}$
sectionL$_{t-4}$	sectionA$_{t-2}$	sectionL$_{t-9}$
gross_capital_formation$_{t-4}$	sectionA$_{t-3}$	sectionL$_{t-10}$
gross_fixed_capital_formation$_{t-4}$	sectionA$_{t-4}$	sectionMN$_{t-3}$
sectionJ$_{t-5}$	sectionA$_{t-5}$	sectionMN$_{t-4}$
sectionGHI$_{t-9}$	sectionA$_{t-6}$	sectionMN$_{t-5}$
sectionGHI$_{t-10}$	sectionA$_{t-8}$	sectionOPQ$_{t-8}$
	sectionBCDE$_{t-2}$	sectionOPQ$_{t-10}$
	sectionBCDE$_{t-4}$	sectionRSTU$_{t-1}$
	sectionBCDE$_{t-9}$	sectionRSTU$_{t-4}$
	sectionF$_{t-2}$	sectionRSTU$_{t-6}$
	sectionF$_{t-5}$	Taxes$_{t-9}$
	sectionF$_{t-7}$	Subsidies$_{t-7}$
	sectionF$_{t-8}$	GG$_{t-8}$
	sectionGHI$_{t-1}$	gross_fixed_capital_formation$_{t-9}$
	sectionGHI$_{t-2}$	export_services$_{t-10}$