KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Multivariate out-of-sample tests for Granger causality

Sarah Gelper and Christophe Croux

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0602

Multivariate Out-of-Sample Tests for Granger Causality

Sarah Gelper and Christophe Croux

Abstract: A time series is said to Granger cause another series if it has incremental predictive power when forecasting it. While Granger causality tests have been studied extensively in the univariate setting, much less is known for the multivariate case. In this paper we propose multivariate out-of-sample tests for Granger causality. The performance of the out-of-sample tests is measured by a simulation study and graphically represented by Size-Power plots. It emerges that the multivariate regression test is the most powerful among the considered possibilities. As a real data application, we investigate whether the consumer confidence index Granger causes retail sales in Germany, France, the Netherlands and Belgium.

Keywords: Consumer Sentiment, Granger Causality, Multivariate Time Series, Out-of-sample Tests

1 Introduction

Suppose we have a multivariate time series y_t of length T containing k components and would like to investigate whether another time series x_t, consisting of l components, Granger causes y_t. The series x_t is said to Granger cause y_t if the past of x_t has additional power in forecasting y_t after controlling for the past of y_t. For this purpose, two models are compared. The full model has y_t as dependent variable and past values of both y_t and x_t as regressors. The restricted model, nested in the full one, has only the past of y_t as regressors. In

this paper, Granger causality tests are carried out by means of an out-of-sample comparison of these two nested models.

The idea of out-of-sample testing is very natural in a Granger causality context. Granger causality questions whether forecasts of one variable can be improved by accounting for the history of another variable. Out-of-sample tests act as if the value of the series at a certain point in time is unknown and predicts this value exclusively on the basis of previous observations. The predicted and the realized value of the series are then compared. If the prediction error using the full model is substantially smaller than the prediction error of the restricted model, it is concluded that the former model has significant better forecasting performance. In contrast to out-of-sample testing, in-sample tests include all observations for model estimation, leading to a risk of overfitting and a too optimistic assessment of the predictive power.

Univariate out-of-sample tests have been proposed by Harvey et al. (1998), Ericsson (1992), and Diebold and Mariano (1995) among others. A thorough discussion of several out-of-sample test statistics for nested models can be found in Clark and McCracken (2001), who compare different tests in a simulation study. Their study, however, is limited to univariate series. By virtue of the growing availability of economic data and the need to understand the relationship between many variables, our aim is to enlarge the applicability of out-of-sample Granger causality tests to the multivariate setting.

Up to our knowledge, no multivariate out-of-sample Granger causality test has yet been proposed. The aim of this study is to construct such tests, examine their performance and apply them to data on consumer confidence and retail sales. Section two describes the multivariate out-of-sample test statistics and Section three compares their performance by the construction of Size-Power plots, see Davidson and McKinnon (1998). The fourth Section briefly discusses the distinction between out-of-sample and in-sample tests. In Section five, the new tests are applied to check whether retail sales are Granger caused by consumer confidence indices in Belgium, Germany, France and the Netherlands. Finally, conclusions are given is Section six.

2 The Test Statistics

Let x_t and y_t be two weakly stationary time series. To test for Granger causality we compare a full and a restricted model. The full model is given by

$$y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \psi_1 x_{t-1} + \ldots + \psi_p x_{t-p} + \varepsilon_{\mathrm{f},t}. \tag{1}$$

Here $\varepsilon_{\mathrm{f},t}$ is a multivariate iid sequence with mean zero and covariance matrix Σ_{f}, and the index t runs from $p+1$ to T. Since y_t is of dimension k and x_t of dimension l, the parameters ϕ_0 up to ϕ_p are $(k \times k)$ matrices, whereas ψ_1 up to ψ_p are rectangular matrices of dimension $(k \times l)$.

The null hypothesis stating that x_t is not Granger causing y_t corresponds to

$$\mathrm{H}_0 : \psi_1 = \psi_2 = \ldots = \psi_p = 0. \tag{2}$$

When this null holds, model (1) reduces to

$$y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_{\mathrm{r},t}, \tag{3}$$

where $\varepsilon_{\mathrm{r},t}$ is a multivariate iid sequence with mean zero and covariance matrix Σ_{r}. Model (3) is called the restricted model and is compared with the full model (1) to test for Granger causality.

An out-of-sample test is conducted in three steps. The first step divides the series y_t in two parts: one containing observations 1 to R and a second part consisting of the following P observations, so that $R + P = T$. The first R observations will always be included for parameter estimation while the last P observations will be forecasted using recursive one-step-ahead forecasts. In the second step, observations $R+1$ up to T are forecasted using a recursive scheme and forecast errors are computed. Both the full (1) and the restricted model (3) are estimated by ordinary least squares using only observations 1 up to R. Based on this estimation, the associated forecasts of observation $R+1$, $\hat{y}_{\mathrm{f},R+1}$ and $\hat{y}_{\mathrm{r},R+1}$, are obtained. Then y_{R+2} is forecasted based on the first $(R+1)$ observations and this procedure is continued recursively up to the end of the series, yielding the series of one-step-ahead forecasts $\hat{y}_{\mathrm{f},t}$ and $\hat{y}_{\mathrm{r},t}$ for t ranging from $R+1$ up to T. The corresponding one-step-ahead forecast errors are then $u_{\mathrm{f},t} = y_t - \hat{y}_{\mathrm{f},t}$ and $u_{\mathrm{r},t} = y_t - \hat{y}_{\mathrm{r},t}$, being vectors of length k. These

vectors are collected into a matrix of dimension $(P \times k)$, where the s-th row contains the vector of one step ahead forecast errors of observation $R + s$. The matrix containing the one-step-ahead forecast errors from the full model will be referred to by u_{f} and from the restricted model by u_{r}. The third step of the out-of-sample testing procedure compares the forecasting performance of the full and the restricted model using u_{r} and u_{f}. We consider three ways of doing so: one based on the comparison of mean squared forecast errors, a regression based test and a test making use of canonical correlations.

One way to test the null hypothesis of no Granger causality is by comparing the mean squared forecast errors (MSFE) of the full and the restricted model. Therefore, the first test statistic is defined as

$$\mathrm{MSFE} = \log\left(\frac{|u_{\mathrm{r}}' u_{\mathrm{r}}|}{|u_{\mathrm{f}}' u_{\mathrm{f}}|}\right), \tag{4}$$

where $|.|$ stands for the determinant of a matrix. If the full model provides better forecasts, the MSFE takes a large value indicating Granger causality. Under the null of no Granger causality, the forecasts from both models are as good and the test statistic is close to zero. In the univariate case, the MSFE test reduces to a Diebold-Mariano test (Diebold and Mariano (1995)), with the squared one-step-ahead forecast error as loss function.

Another way to test for no out-of-sample Granger causality generalizes the approach of Harvey et al. (1998). They describe how no Granger causality corresponds to zero correlation between $u_{\mathrm{r},t}$ and $u_{\mathrm{r},t} - u_{\mathrm{f},t}$. Suppose that one seeks the best forecast combination of $\hat{y}_{\mathrm{r},t}$ and $\hat{y}_{\mathrm{f},t}$:

$$\hat{y}_{\mathrm{new},t} = (1 - \lambda)\hat{y}_{\mathrm{r},t} + \lambda \hat{y}_{\mathrm{f},t}. \tag{5}$$

If λ equals zero the additional regressors in the full model do not improve the prediction and there is no Granger causality, all the information in the full model is also contained in the restricted model. Define e_t as the error of the combined forecast, i.e. $e_t = y_t - \hat{y}_{\mathrm{new},t}$. Using the definitions of $u_{\mathrm{f},t}$ and $u_{\mathrm{r},t}$, it readily follows from (5) that

$$u_{\mathrm{r},t} = \lambda(u_{\mathrm{r},t} - u_{\mathrm{f},t}) + e_t. \tag{6}$$

Testing whether λ equals zero then corresponds to zero correlation between $u_{\mathrm{r},t}$ and $u_{\mathrm{r},t} - u_{\mathrm{f},t}$ and implies no Granger causality.

Since we work in a multivariate setting, zero correlation can be tested by making use of *canonical correlations*. The second multivariate out-of-sample test for Granger causality will therefore be referred to as the canonical correlation test (abbreviated by CC). The null of no Granger causality corresponds to the hypothesis that all canonical correlations between $u_{\mathrm{r},t}$ and $u_{\mathrm{r},t} - u_{\mathrm{f},t}$, denoted by $\rho_1, \rho_2, \ldots, \rho_k$, are zero:

$$\mathrm{H}_0 : \rho_1 = \ldots = \rho_k = 0.$$

This null can be tested by the well known Bartlett test, see for example Johnson and Wichern (2002),

$$\mathrm{CC} = -P \ln \prod_{j=1}^{k} (1 - \hat{\rho}_j^2). \tag{7}$$

The last multivariate out-of-sample test for Granger causality that we consider, is based on directly estimating λ in the regression model (6). Under the null of no Granger causality, we expect the estimated λ to be close to zero. The hypothesis $\lambda = 0$ can be tested by a likelihood ratio test as

$$\mathrm{Reg} = P\big(\log(|u_{\mathrm{r}}' u_{\mathrm{r}}|) - \log(|\hat{\varepsilon}' \hat{\varepsilon}|)\big),$$

where $\hat{\varepsilon}$ is the $(P \times k)$ residual matrix obtained from regression (6).

The multivariate test statistics described here, are a *generalization* of existing univariate tests. When both series are of dimension one, the multivariate regression test is equivalent to the regression test which was proposed by Ericsson (1992), and the canonical correlation test to the encompassing t-test which was proposed by Harvey et al. (1998). The limiting distributions of these two univariate test statistics are non standard and given in the appendix to the paper of Clark and McCracken (2001). They also show that the univariate regression test and correlation test are asymptotically equivalent. The MSFE test in a univariate setting reduces to comparing the sum of squared forecast errors for both models and is equivalent to the original Diebold-Mariano test, see Diebold and Mariano (1995).

In the multivariate setting, the (asymptotic) distribution of the three test statistics is difficult to obtain, since their calculation is not based on original

data but on one step ahead forecast errors. In applications, we propose to compute the critical values of the test statistic by a residual based bootstrap method, as will be discussed in Section five.

3 Size-Power Curves

In this section, we want to compare the performance of the three multivariate out-of-sample Granger causality tests by computing Size-Power curves under fixed alternatives, as described in Davidson and McKinnon (1998). This method allows to compare the power of the tests without knowing the exact distribution of the test statistics under the null of no Granger causality. We select a fixed alternative hypothesis and plot the probability of rejecting the null when the alternative holds versus the rejection probability under the null. The curve is expected to lie above the 45-degree line in the unit square, the larger the distance between the curve and the 45-degree line the better. The most interesting part of the Size-Power plot is the region where the size ranges from zero to 0.2 since in practice a significance level above 20% is never used.

The Size-Power plots are based on simulated data and therefore we need to specify a *Data Generating Process* (DGP) both for the null and for the alternative hypothesis. To generate the data, we first construct x_t according to a VAR(1) model[1]: $x_t = \theta x_{t-1} + \varepsilon_t^x$, where ε_t^x are independent drawings from a multivariate standard normal distribution and the parameter θ (being an $l \times l$ matrix) is chosen randomly subject to the stationarity condition. Subsequently y_t is generated once by model (1) and once by model (3), where the parameters ϕ_0 up to ϕ_p are randomly chosen subject to the stationarity condition. The error terms for generating models (1) and (3) are the same.

A fixed alternative hypothesis needs to be specified by choosing the values of $\psi_1, \ldots, \psi_p$ in model (1). These values should differ enough from zero so that the test statistics are able to detect the difference between the null and the alternative hypothesis. On the other hand, if they are too large, all three curves will be close to the horizontal line at the constant 100% and the Size-Power plot

[1]The conclusions of the simulation study are robust with respect to the choice of the lag length of this VAR-model.

would not be able to distinguish which test performs the best. An alternative hypothesis matching these two requirements in obtained when the elements of ψ_1 up to ψ_p are chosen randomly from a uniform distribution on the interval $]\frac{1}{kp}; \quad \frac{2}{kp}[$. To further specify the DGP, there are several parameters that need to be decided on: the lag length p in (1), the length of the time series T and the number of components in both time series, k and l. In the presented simulation results, representative values of p, T, k and l have been selected.

The Size-Power curves are simulated by carrying out the following three steps:

1. Simulate N=10000 time series of length T under the null and compute for each series the test statistic. Sort the N obtained test statistics from small to large. Denote the i-th value of this ordered sequence by s_i. When s_i is chosen as the critical value, the quantity $(N-i)/(N+1)$ equals the size of the test.

2. Simulate N time series of length T under the fixed alternative hypothesis and compute each time the test statistic. When choosing a certain s_i as critical value, the power of the test is approximated by the fraction of test statistics that exceeds s_i.

3. For each s_i, for i=1 to N, plot the power against the size.

We first consider the case where the specified model is exactly the same as the DGP. Further on, we address the issue of model misspecification, where the specified model does not correspond to the true DGP. Before we present some Size-Power plots, note that two Size-Power curves can only be compared if the alternative hypothesis is the same for both curves.

As a reference setting, we take the length of the time series T equal to 50, $k = l = 2$ and $p = 2$. The resulting Size-Power curves are presented in Figure 1. We clearly see here that the regression test is preferred to both the canonical correlation and the MSFE test: the curve of the regression test is situated further away from the 45-degree line than both other curves. We would like to investigate what happens to the position of the curves when the length of the time series varies and in case of model misspecification.

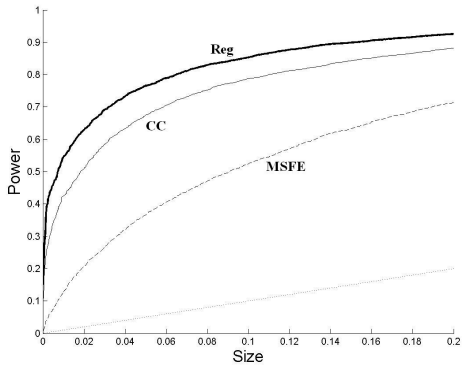
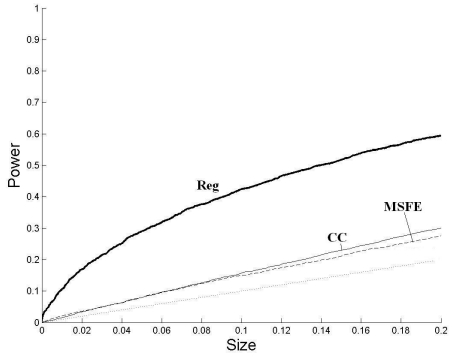


Figure 1: Size-Power plot for T=50, k=2, l=2, p=2. The bold line represents the regression test (Reg), the full line the canonical correlation test (CC) and the dashed line the mean squared forecast error test (MSFE). As a reference, the 45-degree line is presented by the dotted line.

3.1 The effect of the length of the time series

Take now the same setting as in Figure 1, but let the length of the time series vary. We simulate data under the same fixed alternative hypothesis, for T = 20 and 100. The results are presented in Figure 2. First of all, we see that the power of all three tests augments as the length of the time series increases. This is consistent with the general belief that more observations lead to a higher power. For both smaller (T=20) and larger (T = 100) time series, the regression test outperforms the canonical correlation and the MSFE-test. The canonical correlation test is almost as good as the regression test for T = 100. For even longer time series, the canonical correlation and the regression test perform equally well, which is in line with the univariate case in which both tests are shown to be asymptotically equivalent (Clark and McCracken (2001)). But for small time series, the gain in power for the regression test with respect to the CC test and the MSFE test is very prevalent.

(a) $T = 20$



(b) $T = 100$

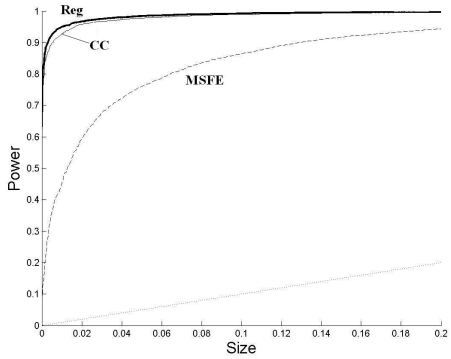


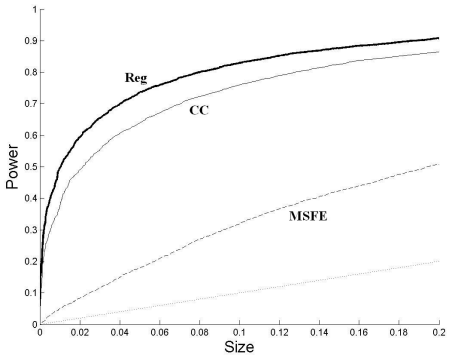
Figure 2: Size-Power plot for varying lengths of T, k=2, l=2, p=2. The bold line represents the regression test (Reg), the full line the canonical correlation test (CC) and the dashed line the mean squared forecast error test (MSFE). As a reference, the 45-degree line is presented by the dotted line.

3.2 Model misspecification

Up until now, the estimated model always perfectly corresponded to the data generating process. When working with real data, however, the model is most probably misspecified because the true DGP is unknown. Therefore, it is interesting to investigate the performance of the out-of-sample tests in the case of model misspecification. As a first form of model misspecification, we consider presence of serial correlation in the error terms of model (1). Computation of the critical value of the test statistics is still based on the assumption of white noise errors, but each component of the error terms of the multivariate DGP follows now an AR(1) process with first order autocorrelation α equal to 0.5 and 0.9 respectively. The higher the value of α, the more we deviate from the specified model. The resulting Size-Power plots are presented in Figure 3, where we choose $T = 50$, $k = 2$, $l = 2$ and p =2 and the same fixed alternative as before. The power of all three tests declines as the degree of serial correlation augments. Though, the ordering of the test statistics does not change; in all settings the regression test is preferred to the canonical correlation test which on its turn is preferred to the MSFE test.

To illustrate the effect of model misspecification under the form of an incorrectly chosen lag length, we consider the matter of selecting a too complex or a too simplistic model respectively. We refer to the lag length of the true DGP by p and of the estimated model by q. Consider first overfitting, estimation of a model which has a larger order than the true model. For computing the test statistics we let the lag length of the estimated model q take the values 3 and 6. The results are presented in Figure 4. When we compare Figure 4 with Figure 1, we see that estimation of a too complex model leads to a drop of power, especially so for the canonical correlation and the MSFE-test. The regression test is still the most powerful of the three considered tests. To get insight in the consequences of underfitting or estimating a too simplistic model, we look at a model of which the true order $p = 6$. The test statistics are based on models having order $q = 2$ and 4. The resulting Size-Power plots are presented in the lower half of Figure 4. The Figure shows that also in this case, the regression test remains most powerful. In summary, the graphs of Figures 4 suggest that

(a) Serial correlation: $\alpha = 0.5$



(b) Serial correlation: $\alpha = 0.9$

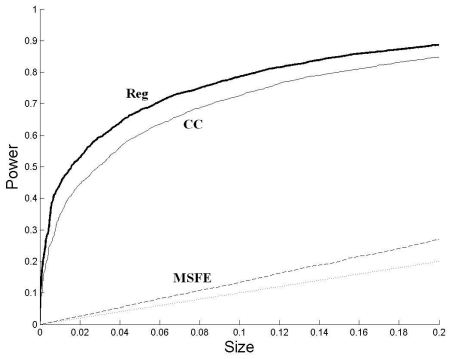


Figure 3: Size-Power plots under model misspecification in the form of varying degrees of autocorrelation in the error terms, with T=50, k=2, l=2 and p=2. The bold line represents the regression test (Reg), the full line the canonical correlation test (CC) and the dashed line the mean squared forecast error test (MSFE). As a reference, the 45-degree line is presented by the dotted line.

the regression test is, among the three tests considered here, the most robust against model misspecification in the form of an incorrectly specified lag length. In every graph presented here, the regression test is more powerful than the canonical correlation test and the MSFE-test. We come to the same conclusion when using other simulation schemes not presented here.

4 Out-of-sample versus in-sample tests

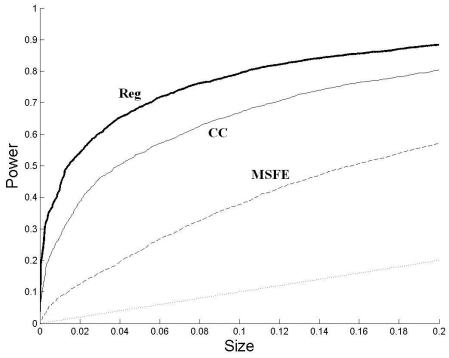
The question whether to use in-sample or out-of-sample model comparison has been studied by various authors. The main concern with in-sample tests is the risk of overfitting the data: if the estimated model is too complex, effects that are found to be significant are suspect to be spurious. The problem of overfitting the data using in-sample procedures is discussed by Clark (2004) for univariate series. By means of a simulation study, he shows that by using of out-of-sample tests, spurious effects can be avoided. On the other hand, Kilian (2002) advocates for the use of in-sample test procedures. He argues that splitting up the data, which is needed for an out-of-sample test, results in a loss of information. Thereby, a deviation from the null hypothesis is less often detected by an out-of-sample test. Especially in small samples out-of-sample testing entails a loss in power. We will remain brief on the issue of in-sample versus out-of-sample testing, since the same arguments given in the univariate case apply to the multivariate problem. We think, however, that in the context of Granger causality tests an out-of-sample approach is more natural, see also Ashley et al. (1980).

In the previous Section, we have seen that among the three considered out-of-sample tests, the multivariate regression test is most powerful. By using the same methodology, namely by drawing Size-Power plots, we compare this test with the standard *in-sample* multivariate likelihood ratio test:

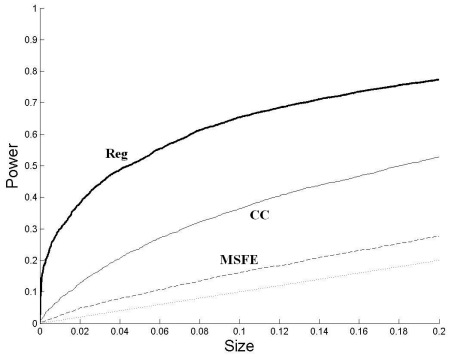
$$\mathrm{LR} = T\big(\log(|\hat{\varepsilon}_{\mathrm{r}}'\hat{\varepsilon}_{\mathrm{r}}|) - \log(|\hat{\varepsilon}_{\mathrm{f}}'\hat{\varepsilon}_{\mathrm{f}}|)\big).$$

The $(T \times k)$ matrices $\hat{\varepsilon}_{\mathrm{f}}$ and $\hat{\varepsilon}_{\mathrm{r}}$ denote the residuals from the estimation of the full and the restricted model respectively, where all observations are included for model estimation. Under the null hypothesis, the LR test statistic has a

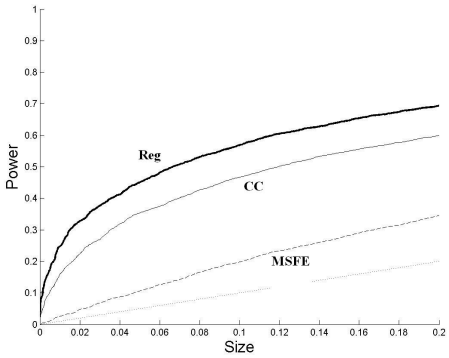
(a) Overfitting: $p = 2$ and $q = 3$



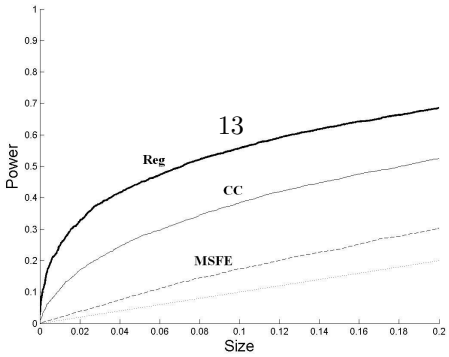
(b) Overfitting: $p = 2$ and $q = 6$



(c) Underfitting: p=6 and q=2



(d) Underfitting: p=6 and q=4



chi-squared distribution with the number of degrees of freedom equal to the number of parameters put equal to zero in expression (2), i.e. pkl.

We find two situations where out-of-sample testing is more powerful than in-sample testing: in case of estimating a too complex model and when dealing with short time series. Corresponding Size-Power curves are pictured in Figure 5. For Figure 5a, the specified model has a larger order than the true DGP, while for Figure 5b the length of the time series is very short, $T = 20$. In both cases, we see that the out-of-sample test is more powerful than the in-sample test. This finding is in line with the previously described concern of overfitting the data, see Clark (2004).

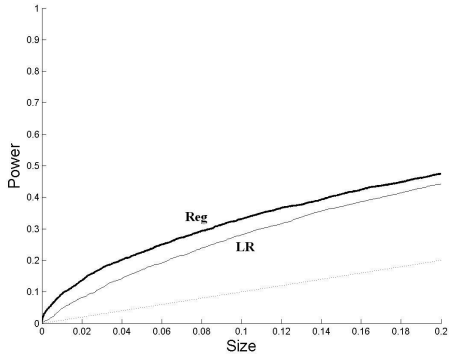
We compared the out-of-sample testing power with the standard in-sample likelihood ratio test. While in-sample testing is very conventional, this study shows that it is not always the most powerful way of testing a null hypothesis of no Granger causality in a multivariate setting. In particular, in the case of a small number of observations and a too complex specified model, the out-of-sample regression procedure outperforms the in-sample likelihood ratio test.

5 Application

The present section studies the predictive power of the consumer confidence index on retail sales in four European countries that are geographically and economically related: Belgium, Germany, France and the Netherlands. The consumer confidence index is often regarded as an indicator of the current economic climate. It is constructed from large scale questionnaires in which the participants are asked what they expect for their future financial situation, and the general economic environment and how these situations have evolved in recent times. One can expect that consumers who are more confident are prepared to spend more. Moreover, if the consumer confidence index Granger causes retail sales, then better forecasts of future sales can be achieved by taking into account the confidence index.

A multivariate testing approach is appropriate here, since it may well be possible that the consumer confidence index of one country Granger causes retail sales in other countries, see also Lemmens et al. (2005). The multivariate test

(a) Estimation of a too complex model:
p=2 and q=6



(b) Short Time Series:
$T = 20$

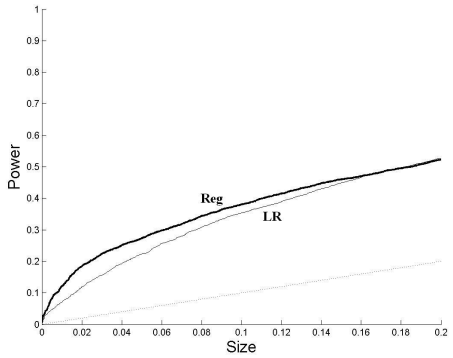


Figure 5: Size-Power plots for the comparison of in-sample and the out-of-sample Granger causality testing, with $k = l = 2$ and p=2. The bold line represents the out-of-sample regression test (Reg), the full line the in-sample likelihood ratio test (LR). As a reference, the 45-degree line is presented by the dotted line.

takes these cross-country relationships into account. The variable of interest is the multivariate time series containing the retail sales figures in the four countries under consideration. The predictor variable is the multivariate time series containing the consumer confidence index in these countries. The data range from January 1985 to December 2002, resulting in 204 observations. To perform the Granger causality tests we discussed earlier, the series need to be stationary. After applying an augmented Dickey-Fuller test[2] to the retail sales data in logarithms, we find a unit root. The series are then taken in differences, yielding the stationary series y_t which represents the monthly growth rate of retail sales. We would like to know whether y_t is Granger caused by the consumer confidence index in the four countries under consideration. Since we also find a unit root[3] in every consumer confidence series, we work with the series in differences, denoted by x_t.

The p-values of the multivariate test statistics, in-sample as well as out-of-sample, are obtained by a bootstrap procedure. As mentioned before, the exact or limiting distribution of the multivariate out-of-sample tests under the null hypothesis of no Granger causality is unknown. Horowitz and Savin (2000) strongly advocate the use of the bootstrap procedure, even when an approximation of the distribution of the statistic under the null can be analytically derived from asymptotic theory. Especially in small samples, asymptotic critical values can lead to serious size-distortion. In this application, critical values will be obtained by a residual based bootstrap, as explained in detail in Davidson and Hinkley (2003). Under the null hypothesis, we assumed the error terms to be iid with mean zero and a certain covariance matrix. This allows for using the residual based bootstrap, and no normality assumption is needed. By estimating the model and resampling from the residuals, 10 000 bootstrap series and associated test statistics are computed. The percentage of bootstrap statistics exceeding the test statistic computed from the observed time series is an approximation of the p-value. An unreported Monte-Carlo study indicates that the size-distortion

[2]The p-values for the null of a unit root in the retail sales data are 0.57 for Belgium, 0.08 for Germany, 0.24 for France and 0.70 for the Netherlands.

[3]The p-values corresponding to the null of a unit root in the consumer confidence series are, 0.08 for Belgium, 0.43 for Germany, 0.06 for France and 0.39 for the Netherlands.

Table 1: Multivariate tests for Granger causality from the consumer confidence towards retail sales. The p-values are presented for the null of no Granger causality.

Test	MSFE	CC	Reg	LR
p-value	0.053	0.069	0.067	0.024

using critical values obtained by such a residual based bootstrap procedure is negligible for $T = 204$, and this for the four test statistics considered in this application, which justifies their use.

The results of the multivariate tests are presented in Table 1. It emerges that all three out-of-sample tests provide rather weak evidence for Granger causality, while the in-sample likelihood ratio test provides stronger evidence. As discussed before, in-sample tests are suspect to detect spurious Granger causality relationships, in particular if the model is misspecified. Therefore we prefer to rely on the out-of-sample procedures, giving only weak evidence of Granger causality. The incremental predictive power of the 4 consumer confidence indices for prediction of the 4 retail series is found to be marginal. The multivariate in-sample Granger causality test is too optimistic. This is in line with a recent study by Garrett et al. (2005), who use multiple univariate in-sample tests for different states in the US, and obtain that the predictive power of consumer confidence for retail sales is limited. Other previous studies, all using univariate tests, report mixed results, see for example Batchelor and Dua (1998), Carroll et al. (1994) and Matsusaka and Sbordone (1995).

6 Conclusion

In this paper, three multivariate out-of-sample testing procedures for Granger causality were proposed and discussed. One is based on directly comparing squared forecast errors, a second is a test based on multivariate regression and the third test relates to canonical correlations. Use of Size-Power plots, a simulation based and computational intensive method, reveals that the regression

based test is the most powerful among the out-of-sample tests. As compared to in-sample testing, the out-of-sample regression test is less sensitive to overfitting. Moreover, out-of-sample testing is more natural in a Granger causality context.

As an application, we looked whether the consumer confidence in Belgium, Germany, France and the Netherlands jointly Granger causes the consumer confidence in these countries. By using multivariate tests, possible cross-relationships between the four countries are taken into account. We found only weak out-of-sample evidence for Granger causality.

References

Ashley, R., Granger, C., Schmalensee, R., 1980. Advertising and aggregate consumption: an analysis of causality. Econometrica 48, 1149–1169.

Batchelor, R., Dua, P., 1998. Improving macro-economic forecasts, the role of consumer confidence. International Journal of Forecasting 14, 71–81.

Carroll, C., Fuhrer, J., Wilcox, D., 1994. Does consumer sentiment forecast household spending? if so, why? The American Economic Review 84, 1397–1408.

Clark, T., 2004. Can out-of-sample forecast comparisons help prevent overfitting? Journal of forecasting 23, 115–139.

Clark, T., McCracken, M., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110.

Davidson, A., Hinkley, C., 2003. Bootstrap methods and their application. Cambridge University press.

Davidson, R., McKinnon, J., 1998. Graphical methods for investigating the size and power of hypothesis tests. The Manchester School 66, 1–26.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. Journal of business and economic statistics 13, 253–263.

Ericsson, N., 1992. Parameter consistency, mean square errors, and measuring forecast performance: an exposition, extensions and illustration. Journal of policy modeling 14, 465–495.

Garrett, T., Hernandez-Murillo, R., Owyang, M., 2005. Does consumer sentiment predict regional consumption? Federal Reserve Bank of St Louis Review 87, 123–135.

Harvey, D., Leybourne, S., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business and Economic Statistics 16, 254–259.

Horowitz, J., Savin, N., 2000. Empirically relevant critical values for hypothesis tests: A bootstrap approach. Journal of Econometrics 95, 375–389.

Johnson, R., Wichern, D., 2002. Applied Multivariate Statistical Analysis. Prentice Hall, New York.

Kilian, L., 2002. In-sample or out-of-sample tests of predictability: which one should we use? European central bank working paper 195.

Lemmens, A., Croux, C., Dekimpe, M., 2005. On the predictive content of production surveys: a pan-european study. International Journal of Forecasting 21, 363–375.

Matsusaka, J., Sbordone, A., 1995. Consumer confidence and economic fluctuations. Economic Inquiry 33, 296–318.