Multivariate out-of-sample tests for Granger causality

Citation for published version (APA):

DOI:
10.1016/j.csda.2006.09.021

Document status and date:
Published: 01/01/2007

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 03. Aug. 2023
Multivariate out-of-sample tests for Granger causality

Sarah Gelper*, Christophe Croux

K.U.Leuven, Faculty of Economics and Applied Economics, Naamsestraat 69, 3000 Leuven, Belgium

Available online 13 October 2006

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. Multivariate out-of-sample tests for Granger causality are proposed and their performance is measured by a simulation study. The results are 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, it is investigated whether the consumer confidence index Granger causes retail sales in Germany, France, the Netherlands and Belgium.

© 2006 Elsevier B.V. All rights reserved.

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 the forecasting performance 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.

* Corresponding author. Tel.: +32 16 326928; fax: +32 16 326732.
E-mail addresses: sarah.gelper@econ.kuleuven.be (S. Gelper), christophe.croux@econ.kuleuven.be (C. Croux).

0167-9473/ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.09.021
By virtue of the growing availability of economic data and the need to understand better the relationship between groups of several variables, the Granger causality concept has been extended to the multivariate setting (e.g. El Himdi and Roy, 1997). Here the aim is to test whether a set of time series, \( x_t \), has incremental predictive power for predicting a second set of time series \( y_t \). In the empirical application we will study, we wonder whether the series of consumer sentiment indices measured in four different countries Granger causes the real sales figures in those four countries. A multivariate approach is appropriate here, since it is well possible that the consumer sentiment index in one particular country has predictive power for the sales figures in the other countries as well. Cross-country relationships are obviously existing, and a multivariate Granger causality test will take these into account. One might argue that alternatively a sequence of \( 4 \times 4 = 16 \) univariate Granger causality tests could be undertaken, but such an approach has several disadvantages. Indeed, using a multivariate procedure will (i) avoid the multiple testing problem and (ii) be more powerful, since it is possible that a single series has only weak Granger causality, but the collection of several series yields a more substantial improvement in predictive power. Multivariate tests for Granger causality (in the in-sample setting) have been studied before and have been shown to be useful. For example, Boudjellaba et al. (1992) find that the vector time series including the two money measures M1 and M2 are Granger causing economic growth in Canada. They also illustrate how using a multivariate approach alters the statistical conclusions as compared to a univariate analysis. Lemmens et al. (2005) use multivariate tests to show that production surveys are Granger causing production levels in the European Union member states, while Ostermarkt and Aaltonen (1999) applied them on asset pricing problems.

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 2 describes the multivariate out-of-sample test statistics and proposes a residual based bootstrap procedure to obtain critical values. Section 3 compares their performance by the construction of size–power plots, as described in Davidson and McKinnon (1998). The fourth section briefly discusses the difference between out-of-sample and in-sample tests. In Section 5, the new procedures are applied to test whether retail sales are Granger caused by consumer confidence indices in Belgium, Germany, France and the Netherlands. Finally, conclusions are given in Section 6.

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

\[
y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \psi_1 x_{t-1} + \cdots + \psi_p x_{t-p} + \epsilon_{t},
\]

Here \( \epsilon_{t} \) is a multivariate iid sequence with mean zero and covariance matrix \( \Sigma_f \), 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 \( \phi_0 \) up to \( \phi_p \) are \( (k \times k) \) matrices, whereas \( \psi_1 \) up to \( \psi_p \) are rectangular matrices of dimension \( (k \times l) \). The null hypothesis stating that \( x_t \) is not Granger causing \( y_t \) corresponds to

\[
H_0 : \psi_1 = \psi_2 = \cdots = \psi_p = 0.
\]

When this null hypothesis holds, model (1) reduces to

\[
y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_{t},
\]

where \( \epsilon_{t} \) is a multivariate iid sequence with mean zero and covariance matrix \( \Sigma_f \). 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 are always included for parameter estimation. In the second step a sequence of \( P \) one-step-ahead forecasts will be made according to a recursive scheme. This scheme starts with estimating the full (1) and the restricted model (3) by ordinary least squares using only observations 1 up to \( R \). Based on this estimation, the associated forecasts of observation \( R + 1 \), \( \hat{y}_{t,R+1} \) and \( \hat{y}_{t,R+1} \), 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 \( \hat{y}_{f,t} \) and \( \hat{y}_{r,t} \) for \( t \) ranging from \( R + 1 \) up to \( T \). Note that the last forecast, \( \hat{y}_{f,T} \) is based on a model estimated from the first \( T - 1 \) observations. Corresponding one-step-ahead forecast errors for \( t = R + 1, \ldots, T \) are computed as \( u_{t,f} = y_t - \hat{y}_{f,t} \) and \( u_{t,r} = y_t - \hat{y}_{r,t} \), being vectors of length \( k \). These vectors are collected into a matrix of dimension \( (P \times k) \), where the \( sth \) row contains the vector of one step ahead forecast errors for observation \( R + s \). The matrix containing the one-step-ahead forecast errors from the full model will be referred to by \( u_t \) and from the restricted model by \( u_t \). The third step of the out-of-sample testing procedure compares the forecasting performance of the full and the restricted model using \( u_t \) and \( u_t \). 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. Throughout the paper we will make the decision how to choose between the two models. 

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. Since we are in a multivariate setting, we will obtain a symmetric positive definite matrix of forecast errors, whose magnitude is measured by taking its determinant. Therefore, the first test statistic is defined as

\[
\text{MSFE} = \log \left( \frac{|u_t' u_t|}{|u_t' u_t|} \right),
\]

where \(|.|\) stands for the determinant of a matrix. Alternatively, one could take the trace instead of the determinant, but this would result in a test statistic being (i) less powerful, since correlations are not taken into account by the trace operator and (ii) not invariant if the units of measurement of the variables to predict are changing; so if the first component of \( y_t \) is multiplied by 10, while the other components remain unchanged, the value of the test statistic will change if we use the trace, but not if we use the determinant. 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 comparable and the test statistic is close to zero.

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 \( u_{t,f} \) and \( u_{t,r} - u_{t,f} \). Suppose that one seeks the best forecast combination of \( \hat{y}_{f,t} \) and \( \hat{y}_{r,t} \)

\[
\hat{y}_{\text{new},t} = (1 - \lambda) \hat{y}_{r,t} + \lambda \hat{y}_{f,t}.
\]

If \( \lambda \) 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. \( e_t = y_t - \hat{y}_{\text{new},t} \). Using the definitions of \( u_{t,r} \) and \( u_{t,f} \), it readily follows from (5) that

\[
u_{t,r} = \lambda (u_{t,r} - u_{t,f}) + e_t.
\]

Testing whether \( \lambda \) equals zero then corresponds to zero correlation between \( u_{t,r} \) and \( u_{t,r} - u_{t,f} \) 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 \( u_{t,r} \) and \( u_{t,r} - u_{t,f} \), denoted by \( \rho_1, \rho_2, \ldots, \rho_k \), are zero

\[
H_0 : \rho_1 = \cdots = \rho_k = 0.
\]

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

\[
\text{CC} = -P \ln \prod_{j=1}^{k} \left( 1 - \rho_j^2 \right).
\]
The last multivariate out-of-sample test for Granger causality that we consider, is based on directly estimating \( \lambda \) in the regression model (6). Under the null of no Granger causality, we expect the estimated \( \hat{\lambda} \) to be close to zero. The hypothesis \( \lambda = 0 \) can be tested by a likelihood ratio test as

\[
\text{Reg} = P(\log(|u'_{t}u_{t}|) - \log(|\hat{\mu}^2|)),
\]

where \( \hat{\lambda} \) 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 similar to the Diebold–Mariano test, see Diebold and Mariano (1995), with the squared one-step-ahead forecast error as loss function.

The (asymptotic) distribution of the three multivariate test statistics under the null is difficult to obtain, since their calculation is not based on original data but uses one step ahead forecast errors. Asymptotic inference in the univariate case has been studied by West (1996) for the MSFE test and by Clark and McCracken (2001) for the regression and correlation based tests. We do not elaborate on the analytics of the distribution of the tests, but propose to compute approximate critical values and \( p \)-values by a residual based bootstrap method. 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.

We now outline how the residual based bootstrap for obtaining the critical values or \( p \)-values is applied; a more detailed treatment can be found in Davidson and Hinkley (2003):

1. Estimate the model under the null hypothesis, i.e. model (3), using the series \( y_1, y_2, \ldots, y_T \). Compute the sequence of residuals \( r_1, r_2, \ldots, r_T \). Compute the value of the test statistic, denoted here by \( s_0 \).
2. Generate \( Nb = 1000 \) new time series \( y_1^s, y_2^s, \ldots, y_T^s \) according to model (3), with the unknown parameters replaced by their estimates, and the error terms replaced by a bootstrap sample (so resampling with replacement) from \( r_1, r_2, \ldots, r_T \). Compute for each of the \( Nb \) series the value of the test statistic, resulting in \( s_1^s, \ldots, s_{Nb}^s \). For computing the test statistics one also uses the values of \( x_t \), which are kept fix.
3. The percentage of bootstrap replicates \( s_1^s, \ldots, s_{Nb}^s \) exceeding \( s_0 \) is an approximation of the \( p \)-value. The \( \alpha \)-quantile of the bootstrap replicates serves a critical value of the test at level \( \alpha \).

The bootstrap distribution of the residuals is expected to be close to the true distribution of the error terms, in particular for larger sample sizes. If we would sample form the true distribution, then the Monte-Carlo procedure is exact for \( Nb \) tending to infinity. The residual based bootstrap relies on the assumptions of iid error terms and a correctly specified model. Note that no normality assumption is needed. Moreover, the residual bootstrap allows for instantaneous cross correlations between the components of the error term, since the resampling from the multivariate residuals preserves possible correlations. Kilian (1999) performed a more in depth examination of the residual based bootstrap method and its validity. In the next section, it is shown by means of a simulation study that the size distortion is negligible for not too small values of \( T \).

3. Size-power curves

This section studies the size and power of the three multivariate out-of-sample Granger causality tests by means of a simulation study. First we specify the data generating process (DGP) under the null hypothesis. To generate the data, we construct the predictor variables \( x_t \) according to a VAR(1) model: \( x_t = \theta x_{t-1} + e_t \), where \( e_t \) are independent drawings from a multivariate standard normal distribution and the parameter \( \theta \) (being an \( l \times l \) matrix) is chosen arbitrarily subject to the stationarity condition. The conclusions of the simulation study are robust with respect to the choice of the lag length of this VAR-model. Subsequently, \( y_t \) is generated according to model (3), where the parameters \( \phi_0 \) up to \( \phi_P \) (being \( k \times k \) matrices) are arbitrarily selected subject to the stationarity condition, and remain fixed during the
Table 1
Simulated size of the regression test (Reg.), the canonical correlation test (CC) and the mean squared forecast error test (MSFE), for four different sampling schemes

<table>
<thead>
<tr>
<th>Sampling Scheme</th>
<th>Reg.</th>
<th>CC</th>
<th>MSFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I) T = 50, k = 2, l = 2, p = 2</td>
<td>α = 5%</td>
<td>0.050</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>α = 10%</td>
<td>0.110</td>
<td>0.103</td>
</tr>
<tr>
<td>(II) T = 20, k = 2, l = 2, p = 2</td>
<td>α = 5%</td>
<td>0.039</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>α = 10%</td>
<td>0.087</td>
<td>0.109</td>
</tr>
<tr>
<td>(III) T = 100, k = 2, l = 2, p = 2</td>
<td>α = 5%</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>α = 10%</td>
<td>0.109</td>
<td>0.100</td>
</tr>
<tr>
<td>(IV) T = 204, k = 4, l = 4, p = 2</td>
<td>α = 5%</td>
<td>0.044</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>α = 10%</td>
<td>0.095</td>
<td>0.093</td>
</tr>
</tbody>
</table>

simulation runs. For the covariance matrix of the normally distributed innovations we take the identity matrix, which is without loss of generality due to the equivariance of the different tests. To further specify the DGP, there are several parameters that need to be decided on: the lag length \( p \), 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 performance ranking of the different estimators was hardly affected when taking other values of the DGP parameters.

To illustrate the appropriateness of the residual based bootstrap, we examine the size of the tests using bootstrap critical values. We simulate 1000 series under the above specified DGP, satisfying the null hypothesis, and compute the fraction of test statistics exceeding the bootstrap critical value (note that both the test statistic and the critical value are different in every simulation run), both at 5% and the 10% level. The reported results can be found in Table 1 for four different simulation schemes. The last setting we consider, \( T = 204, k = 4, l = 4, p = 2 \), corresponds to the empirical application we will discuss in Section 5. Standard errors around the reported result are about 0.6% for \( \alpha = 0.05 \), and 0.1% for \( \alpha = 0.1 \). Recall that \( P = R = T/2 \), where we acknowledge that the choice of \( R \) may have an impact on the statistical behavior of the different tests (see Clark and McCracken, 2001; Clark, 2004).

One can see from Table 1 that the empirical sizes are relative close to the required values of \( \alpha \). For \( T = 20 \) the size distortion is most severe, but still within bounds, given that this setting is corresponding to very short time series of only 20 observations. The MSFE suffers most from size distortion. We conclude that the residual based bootstrap method for obtaining critical values is sufficiently accurate for the different models considered here, in particular for the REG and the CC test.

The power of the three test procedures will now be compared by simulating size–power curves under fixed alternatives, as described in Davidson and McKinnon (1998). Size–power plots allow to compare the power of the different tests and the CC test. For another recent application of size–power plots in the statistical literature, see Siani and de Peretti (2006). Since all three out-of-sample tests are sufficiently accurately sized, the size–power plots will allow for a fair comparison of the power of the different tests.

A fixed alternative hypothesis needs to be specified by selecting values of \( \psi_1, \ldots, \psi_p \) in model (1). Here, the \( \psi_1, \ldots, \psi_p \) are drawn from a uniform distribution on the interval \( [1/kp; 2/kp] \), and remain fixed afterwards. Typically, these values yield size–power plots allowing us to distinguish the power of the different tests. The size–power curves are then 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.
The size–power curve is expected to lie above the 45° line in the unit square, the larger the distance between the curve and the 45° 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. As a reference setting, we take the length of the time series \( T \) equal to 50, \( k = l = 2 \) and \( p = 2 \), corresponding to scheme (I) in Table 1. The resulting size–power curves are presented in Fig. 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° 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 what happens in case of model misspecification.

### 3.1. The effect of the length of the time series

Take now the same setting as in Fig. 1, but let the length of the time series vary. We simulate data under the same fixed alternative hypothesis for \( T = 20 \) and 100, corresponding to schemes (II) and (III) in Table 1. The results are presented in Fig. 2. The size–power plot for Scheme (IV), corresponding to the empirical application of Section 5, is very similar to Scheme (III) and not reported here. 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.

### 3.2. Model misspecification

Up until now, the estimated model always corresponded to the true 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). Each component of the error terms of the multivariate DGP follows now an AR(1) process with first order autocorrelation \( \alpha \) equal to 0.5 and 0.9,
respectively. The higher the value of $z$, the more we deviate from the specified model. Computation of the critical value of the test statistics is still based on the assumption of white noise errors. The resulting size–power plots are presented in Fig. 3, for simulation setting (I), $T = 50, k = 2, l = 2$ 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 Fig. 4. When we compare Fig. 4 with Fig. 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 Fig. 4. The graph shows that also in this case, the regression test remains most powerful. In summary, the graphs in Fig. 4 suggest that 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 the use of out-of-sample tests can avoid spurious effects. On the other hand, Inoue and Kilian (2004) advocate 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 could less often be detected by an out-of-sample test. Especially in small samples, out-of-sample testing may entail 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:

\[
LR = T \log(|\hat{\varepsilon}_r^T \hat{\varepsilon}_r|) - \log(|\hat{\varepsilon}_f^T \hat{\varepsilon}_f|).
\]

The \((T \times k)\) matrices \(\hat{\varepsilon}_f\) and \(\hat{\varepsilon}_r\) denote the residuals from the estimation of the full and the restricted model, respectively, where all observations are included for model estimation. Note that the in-sample test is based on \(T\) “forecast errors”, while the out-of-sample tests are constructed from only \(P\) recursively obtained one-step-ahead forecast errors. The latter are based on a sequence of model estimates using up to \(T - 1\) observations. Under the null hypothesis, the LR test statistic has a 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 Fig. 5. For Fig. 5a, the specified model has a larger order than the true DGP, while for Fig. 5b the length of the time series is very short, \(T = 20\). In both cases, we see that the out-of-sample regression 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).

In this section, 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 may outperform the in-sample likelihood ratio test. Of course, for time series with a large enough time span and under a correctly specified model, the likelihood ratio test will always be the more powerful.
Table 2
Multivariate tests for Granger causality from the consumer confidence towards retail sales

<table>
<thead>
<tr>
<th>Test</th>
<th>MSFE</th>
<th>CC</th>
<th>Reg.</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.053</td>
<td>0.069</td>
<td>0.067</td>
<td>0.024</td>
</tr>
</tbody>
</table>

The p-values are presented for the null of no Granger causality.

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, for 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 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 to the retail sales data in logarithms, we find a unit root. The series are then taken in differences, yielding the four-dimensional 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 in every consumer confidence series, we work with the series in differences, resulting in \( x_t \). Before estimating model (1), the lag length \( p \) needs to be specified. For this we use the Bayes information criterion (BIC), resulting in \( p = 2 \). The different test statistics are then computed, and the corresponding p-values are obtained by a residual bootstrap procedure, as described in Section 2.

The results of the multivariate tests are presented in Table 2. 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 four consumer confidence indices for prediction of the four 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.

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, the second is based on multivariate regression and the third one 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 retail sales in these countries. By using a multivariate approach, possible cross-relationships between the four countries are taken into account. In contrast to the over-optimistic in-sample approach, we found only weak out-of-sample evidence for Granger causality.
Given the increasing availability of multiple time series, we think that there is an increasing need for multivariate approaches in time series analysis (see Prado et al., 2006 for another recent contribution to multivariate time series). While the present paper is only testing whether multivariate Granger causality is present or not, it may also be useful to measure the strength of the Granger causality, and to decompose the predictive power over different forecasting horizons. In the univariate case, several approaches are existing for a more in-depth analysis of the Granger causality (e.g. Gelper et al., 2006). A similar decomposition for the multivariate out-of-sample case is part of our future research plans.

Acknowledgments

We would like to thank the editor and two referees for constructive comments. We also thank the ‘Fonds voor Wetenschappelijk Onderzoek’ which funded this research (Contract Number G.0594.05).

References