SFB 649 Discussion Paper 2008-073

Testing directional forecast value in the presence of serial correlation

Oliver Blaskowitz*
Helmut Herwartz**

* Humboldt–Universität zu Berlin, Germany
** Christian–Albrechts–Universität zu Kiel, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Testing directional forecast value in the presence of serial correlation *

Oliver Blaskowitz
Humboldt–Universität zu Berlin
Institute of Statistics and Econometrics
Phone: +49 – 30 – 2093 – 5705
Email: blaskowitz@wiwi.hu–berlin.de

Helmut Herwartz
Christian–Albrechts–Universität zu Kiel
Institute of Statistics and Econometrics
Phone: +49 – 431 – 880 – 2417
Email: herwartz@stat–econ.uni–kiel.de

Abstract

Common approaches to test for the economic value of directional forecasts are based on the classical χ^2–test for independence, Fisher's exact test or the Pesaran and Timmerman (1992) test for market timing. These tests are asymptotically valid for serially independent observations. Yet, in the presence of serial correlation they are markedly oversized as confirmed in a simulation study. We summarize serial correlation robust test procedures and propose a bootstrap approach. By means of a Monte Carlo study we illustrate the relative merits of the latter. Two empirical applications demonstrate the relevance to account for serial correlation in economic time series when testing for the value of directional forecasts.

Keywords: Directional forecasts, directional accuracy, forecast evaluation, testing independence, contingency tables, bootstrap.
JEL classification: C32, C52, C53, E17, E27, E47, F17, F37, F47, G11, G17

*This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

1 Introduction

Forecasts are produced in numerous areas as they are important tools for decision making. The implication of a decision based on a forecast can be evaluated by means of the (expected) gain/loss associated with the decision. A commonly used loss function for quantitative forecasts is the quadratic loss of the forecast error. Yet, the squared forecast error provides only a partial assessment of economic forecasts. Diebold and Mariano (1995) point out that in light of the variety of economic decision problems relying on forecasts, statistical loss functions such as quadratic loss need not necessarily conform to economic loss functions. Granger and Pesaran (2000) discuss relationships between statistical and economic measures of forecast accuracy and stress that the choice of the evaluation measure should be related to the objectives of the forecast user. Assessing the directional accuracy (DA) of predicted directions may provide valuable insights into forecast evaluation. Lai (1990) emphasizes that an investor can still gain profits even with statistically biased forecasts if they are on the correct side of the price change more often than not. Leitch and Tanner (1995) find that DA is highly correlated with profits in an interest rate setting. As standard measures such as mean squared/absolute forecast error (MSFE, resp. MAFE) are less correlated with profits, they conclude that DA is a better measure of forecast accuracy for profit maximizing firms. Ash, Smith and Heravi (1998) note that qualitative statements such as the economy is expanding or the economy is contracting in the near future are important pre–requisites for an appropriate implementation of monetary and fiscal policy. Öller and Barot (2000) point out that DA is of interest for central banks. A forecast of increased inflation (above target) would prompt central banks to raise interest rates.

An approach to assess directional forecasts which is linked but not equivalent to the loss functional approach is based on Merton (1981). He proposes an equilibrium theory for the economic value of market timing skills and provides a statistic to measure the value. Cicarelli (1982) uses the statistical measure to analyze turning point errors. Havenner and Modjtahedi (1988), Breen, Glosten and Jagannathan (1989), Schnader and Stekler (1990), Lai (1990) and Stekler (1994) were among the first to apply Merton's theory to evaluate the economic value of directional forecasts. More recent applications include, inter alia, Ash, Smith and Heravi (1998), Mills and Pepper (1999), Öller and Barot (2000), Pons (2001), Easaw, Garratt and Heravi (2005) and Ashiya (2003, 2006). Considering realized and forecasted directions

as binary variables, Merton's theory implies that directional forecasts have no value if the directional outcomes and forecasts are independent. Henriksson and Merton (1981) propose statistical procedures for evaluating forecasting skills that are in fact related to Fisher's (1934) exact test for testing whether two binary variables are independent. Similarly, the classical asymptotic χ^2–test for independence and the asymptotic test for market timing introduced by Pesaran and Timmerman (1992, PT92 henceforth) can be used for testing the economic value of directional forecasts. Yet, these tests are derived under the assumption of serial independence. As we outline later, they are seriously oversized in the presence of serially correlated forecasted resp. realized directions.

Recently, Pesaran and Timmerman (2008, PT08 henceforth) have introduced statistics for testing dependence among serially correlated multi–category variables which can be used to test for the economic value of directional forecasts in the more realistic situation of serial correlation. However, their test procedures reveal some small sample size distortions in a Monte Carlo simulation study. In this paper, we summarize and analyze the size and power properties of a battery of tests for the economic value of directional forecasts in the presence of serial correlation. Furthermore, we propose a bootstrap test procedure to reduce size distortions in small samples. We show in a simulation study that the bootstrap test is robust to serial correlation and has appealing power properties. Our approach can be put in a more general framework, i.e. testing dependence of two binary variables in the presence of serial correlation. Moreover, it can be easily extended to multi–categorical data.

The remainder of the paper is organized as follows. We briefly review Merton's approach in the next Section. In Section 3 existing test procedures and the bootstrap approach are summarized. Section 4 documents a Monte Carlo study to analyze size and power properties of the tests. Section 5 provides two empirical applications and Section 6 concludes.

2 Merton's framework for evaluating directional forecasts

Merton (1981) proposes an equilibrium theory for the value of market timing skills. In the context of evaluating directional forecasts for a variable of interest Y_t, let realized upward resp. downward movements in Y_t be denoted by $\widetilde{Y}_t = 1$, respectively, $\widetilde{Y}_t = 0$. Forecasted

upward resp. downward movements are denoted by $\widetilde{X}_t = 1$ resp. $\widetilde{X}_t = 0$. It is assumed that forecasts $\widetilde{X}_t$ are determined using only information up to time $t-1$. A directional forecast has no value in the sense of Merton (1981) if and only if

$$\mathbb{P}[\widetilde{X}_t = 1|\widetilde{Y}_t = 1] + \mathbb{P}[\widetilde{X}_t = 0|\widetilde{Y}_t = 0] = 1 \ . \tag{2.1}$$

In (2.1) $\mathbb{P}[\widetilde{X}_t = 1|\widetilde{Y}_t = 1]$ ($\mathbb{P}[\widetilde{X}_t = 0|\widetilde{Y}_t = 0]$) denote the conditional probability of a correct forecast of an upward (downward) movement. To alleviate notation, we define $HM = \mathbb{P}[\widetilde{X}_t = 1|\widetilde{Y}_t = 1] + \mathbb{P}[\widetilde{X}_t = 0|\widetilde{Y}_t = 0]$. For example, if $\widetilde{X}_t$ and $\widetilde{Y}_t$ are independent then $\mathbb{P}[\widetilde{X}_t = 1|\widetilde{Y}_t = 1] = \mathbb{P}[\widetilde{X}_t = 1]$ and $\mathbb{P}[\widetilde{X}_t = 0|\widetilde{Y}_t = 0] = \mathbb{P}[\widetilde{X}_t = 0]$. Consequently, $HM = 1$ and such directional forecasts have no value. In particular, naively forecasting only one direction, say $\widetilde{X}_t = 1 \ \forall t$, has no value.

Moreover, Merton (1981) points out that directional forecasts have positive value if and only if

$$HM > 1$$

and that the larger HM the larger the value. Noteworthy, it can be shown that

$$HM - 1 = \frac{\text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right)}{\mathbb{V}\left[\widetilde{Y}_t\right]} \ ,$$

where $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t) = \mathbb{P}[\widetilde{X}_t = 1, \widetilde{Y}_t = 1] - \mathbb{P}[\widetilde{X}_t = 1]\mathbb{P}[\widetilde{Y}_t = 1]$ and $\mathbb{V}[\widetilde{Y}_t] = \mathbb{P}[\widetilde{Y}_t = 1] - \mathbb{P}[\widetilde{Y}_t = 1]^2$ denote the covariance between $\widetilde{X}_t$ and $\widetilde{Y}_t$ and the variance of $\widetilde{Y}_t$, respectively. Hence, the value of the forecasts can be assessed by means of the covariability of realized and forecasted directions. In particular, directional forecasts have (i) no value if and only if $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t) = 0$ and (ii) have value if and only if $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t) > 0$. Moreover, (iii) for a given process Y_t and hence $\widetilde{Y}_t$ (resp. $\mathbb{V}[\widetilde{Y}_t]$), it holds that the larger $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t)$ the larger the value.

Furthermore, maximizing $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t)$ is not equivalent to maximizing the probability of a correct directional forecast $\mathbb{P}[\widetilde{Z}_t = 1]$, where $\widetilde{Z}_t = I(\widetilde{X}_t = \widetilde{Y}_t)$ and $I(\bullet)$ denotes an indicator function. From the relationship

$$\text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) = \frac{1}{2}\mathbb{P}[\widetilde{Z}_t = 1] + \mathbb{P}[\widetilde{X}_t = 1]\left(\frac{1}{2} - \mathbb{P}[\widetilde{Y}_t = 1]\right) + \frac{1}{2}\left(\mathbb{P}[\widetilde{Y}_t = 1] - 1\right)$$

it can be seen that the correspondence between $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t)$ and $\mathbb{P}[\widetilde{Z}_t = 1]$ is not monotone.

Consequently, for a given process Y_t, if the probability of a correct forecast $\mathbb{P}[\widetilde{Z}_t = 1]$ increases and the probability of an upward movement forecast $\mathbb{P}[\widetilde{X}_t = 1]$ changes, then

$$\Delta \text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) = \frac{1}{2}\Delta\mathbb{P}[\widetilde{Z}_t = 1] + \Delta\mathbb{P}[\widetilde{X}_t = 1]\left(\frac{1}{2} - \mathbb{P}[\widetilde{Y}_t = 1]\right) ,$$

with Δ denoting the total difference operator. Whether $\text{Cov}(\widetilde{X}_t, \widetilde{Y}_t)$ increases depends on the signs and magnitudes of $\Delta\mathbb{P}[\widetilde{X}_t = 1]$ and $\frac{1}{2} - \mathbb{P}[\widetilde{Y}_t = 1]$.

Moreover, the loss functional approach as defined below is not equivalent to the Merton approach. Frequently, loss functions to assess DA are defined as:

$$L_t = \begin{cases} a & \text{if } \widetilde{Z}_t = 1 \\ b & \text{if } \widetilde{Z}_t = 0, \end{cases}$$

where $(a,\ b) \neq 0$. Examples include Leitch and Tanner (1995), Greer (2005), Blaskowitz and Herwartz (2008) where $(a,\ b) = (1, -1)$ or Swanson and White (1995, 1997a,b), Gradojevic and Yang (2006) and Diebold (2007) with $(a, b) = (1, 0)$. Hence, a correct directional forecast implies a loss of a (this is rather a gain if $a > 0$) and an incorrect directional forecast implies a loss of b. In this case the expected DA is given by

$$\mathbb{E}[L_t] = (a - b)\mathbb{P}[\widetilde{Z}_t = 1] + b \ .$$

Consequently, maximizing expected DA is equivalent to maximizing the probability of a correct directional forecast (if $a > b$). See Pesaran and Skouras (2002) for a link between the HM statistic and a loss functional approach in a decision–based forecast evaluation framework. For test procedures using loss functions in the presence of serial correlation see, inter alia, Diebold and Mariano (1995) and West (2006).

Note that the value of directional forecasts in the sense of Merton does not take the magnitudes of realized and forecasted changes into account. Hence, the Merton framework is also different from the directional accuracy test proposed in Anatolyev and Gerko (2005) and from the notion of directional forecast value considered in Blaskowitz and Herwartz (2008).

3 Testing for zero covariance

In this section, we first summarize some classical procedures to test for zero covariance between two categorical random variables when there is no serial dependence. Second, we

describe tests for zero covariance in the presence of serial correlation and propose some bootstrap procedures to account for small sample size distortions. We consider tests of the null hypothesis

$$H_0 : \text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) = 0 \ .$$

Notably, if $\widetilde{X}_t$ and $\widetilde{Y}_t$ are Bernoulli variables

$$\text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) = 0 \ \Leftrightarrow \ \widetilde{X}_t \text{ and } \widetilde{Y}_t \text{ are independent} \ .$$

3.1 Testing for zero covariance under serial independence

In the framework outlined above it is straightforward to use 2×2 contingency tables whenever $\widetilde{X}_t$ and $\widetilde{Y}_t$ are serially independent. Testing H_0 can be accomplished using the asymptotic χ^2–test for independence. For small sample sizes Fisher's test (Fisher 1934) based on the hypergeometric distribution is exact and the uniformly most powerful unbiased (UMPU) test for H_0 when the marginals are fixed. If the latter condition does not hold, Fisher's test is no longer exact in finite samples but is asymptotically equivalent to the χ^2–test, see Agresti (1992) for a survey of exact inference for contingency tables.

PT92 proposed a test based on the difference of $\mathbb{P}[\widetilde{Z}_t = 1]$ under dependence and the probability of $\widetilde{Z}_t = 1$ under independence of $\widetilde{Y}_t$ and $\widetilde{X}_t$. In the former case it holds

$$\mathbb{P}[\widetilde{Z}_t = 1] = \mathbb{P}[\widetilde{Y}_t = 1, \widetilde{X}_t = 1] + \mathbb{P}[\widetilde{Y}_t = 0, \widetilde{X}_t = 0] \ .$$

If $\widetilde{Y}_t$ and $\widetilde{X}_t$ are independently distributed the probability of $\widetilde{Z}_t = 1$ is given by

$$\mathbb{P}_{indep}[\widetilde{Z}_t = 1] = \mathbb{P}[\widetilde{Y}_t = 1]\mathbb{P}[\widetilde{X}_t = 1] + \mathbb{P}[\widetilde{Y}_t = 0]\mathbb{P}[\widetilde{X}_t = 0] \ .$$

Hence, the test proposed by PT92 is based on

$$PT = \mathbb{P}[\widetilde{Z}_t = 1] - \mathbb{P}_{indep}[\widetilde{Z}_t = 1] = 2\text{Cov}\left(\widetilde{Y}_t, \widetilde{X}_t\right) \ .$$

Consequently, $\text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) = 0$ if and only if $PT = 0$. Under the assumption of serial independence of $\widetilde{Y}_t$ resp. $\widetilde{X}_t$ and using a Hausman–type argument their proposed scaled test statistic is asymptotically Gaussian. Moreover, this test is asymptotically equivalent to the χ^2–test when two binary variables are considered. Granger and Pesaran (2000) and Pesaran and Skouras (2002) also derive a relationship between the HM statistic and the statistic proposed in PT92.

The three test procedures described above are frequently used within the context of directional forecast evaluation. The χ^2–approach is applied, inter alia, by Schnader and Stekler (1990), Artis (1996), Kolb and Stekler (1996), Swanson and White (1997a, 1997b), Ash, Smith and Heravi (1998), Mills and Pepper (1999), Öller and Barot (2000), Pons (2000, 2001), Easaw, Garratt and Heravi (2005) and Greer (2003, 2005). Applications of Fisher's test to analyse the value of directional forecasts include, among others, Havenner and Modjtahedi (1988), Lai (1990), Kuan and Liu (1995), Swanson and White (1995, 1997a, 1997b), Gençay (1998), Ash, Smith and Heravi (1998), Joutz and Stekler (1998, 2000), Easaw, Garratt and Heravi (2005) and Ashiya (2003, 2006). The test statistic proposed by PT92 is used, for example, by Pesaran and Timmerman (1995), Kuan and Liu (1995), Ash, Smith and Heravi (1998), Gençay (1998), Mills and Pepper (1999), Pons (2001), Schneider and Spitzer (2004) and Easaw, Garratt and Heravi (2005).

Another approach to test for zero covariance, which is useful when considering serial correlation over time, is given by the bivariate regression model

$$\widetilde{X}_t = \alpha + \beta \widetilde{Y}_t + \varepsilon_t \ , \tag{3.1}$$

where ε_t is a discrete zero mean random error. Note that for the population coefficient it holds $\beta = \mathrm{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) / \mathbb{V}[\widetilde{Y}_t]$. Hence, testing H_0 amounts to standard significance tests for β in a linear regression model. Note, that we regard the regression model merely as a tool for testing purposes only. In our context the model in (3.1) does not have a 'causal' or 'economic' interpretation in the usual sense. Hence, it is also conceivable to regress $\widetilde{Y}_t$ on $\widetilde{X}_t$. These two approaches are asymptotically equivalent under the null hypothesis and differ only in terms of power (Anatolyev, 2006).

Moreover, consider the logistic regression model

$$\widetilde{X}_t = \frac{\exp\left(\alpha + \beta \widetilde{Y}_t\right)}{1 + \exp\left(\alpha + \beta \widetilde{Y}_t\right)} + \varepsilon_t \ ,$$

where ε_t is a discrete zero mean disturbance term. In this model with two binary variables it can be shown that

$$\frac{\mathrm{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right)}{\mathbb{V}\left[\widetilde{Y}_t\right]} = \left(e^{\beta} - 1\right) \mathbb{P}[\widetilde{X}_t = 1 | \widetilde{Y}_t = 0] \mathbb{P}[\widetilde{X}_t = 0 | \widetilde{Y}_t = 1]$$

(Cox and Hinkley, 1974). Again, it follows that $\text{Cov}\left(\widetilde{X}_t, \widetilde{Y}_t\right) = 0$ if and only if $\beta = 0$. Standard maximum likelihood estimation and likelihood ratio (LR) tests can be applied. The small sample distribution of the LR statistic is generally unknown but for Bernoulli variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ the small sample LR test for $\beta = 0$ corresponds to Fisher's exact test (Cumby and Modest, 1987).

3.2 Testing for zero covariance in the presence of serial correlation

When there is serial dependence, the tests described above are no longer suitable. Bartlett (1951) and Patankar (1954) were among the first to show that for (Markov) dependent data the usual Pearson statistic for testing goodness of fit need not have common asymptotic χ^2–distribution. Within the framework of 2×2 contingency tables, Altham (1979) reports an inflated χ^2–statistic, $X^2_{I,T}$, when analyzing relationships between categorical variables observed over time and provides upper and lower bounds for the appropriate test statistic. Tavaré and Altham (1983) show that the classical χ^2 test statistic for independence is either inflated or deflated if $\widetilde{X}_t$ resp. $\widetilde{Y}_t$ are two–state Markov chains. For a general $r \times c$ contingency table Holt, Scott and Ewings (1980) and Tavaré (1983) establish that the asymptotic distribution of $X^2_{I,T}$ depends on unknown nuisance parameters under the null hypothesis if (in this case the multi–categorical variables) $\widetilde{X}_t$ resp. $\widetilde{Y}_t$ are arbitrary (but positive recurrent) Markov chains. Noteworthy, Tavaré (1983) also demonstrates that $X^2_{I,T}$ is still asymptotically distributed as χ^2 with $(r-1)(c-1)$ degrees of freedom when one process, say $\widetilde{X}_t$, is serially independent. Yet, if $\widetilde{Y}_t$ are directions of a serially correlated economic time series and $\widetilde{X}_t$ are reasonable directional forecasts of $\widetilde{Y}_t$ then both processes most likely exhibit serial correlation.

Furthermore, PT08 show in a simulation experiment that the test for market timing proposed in PT92 is seriously oversized in the presence of serial dependence. Finally, it is well known that coefficient tests in a regression model are size distorted if serial correlation is not taken into account. In the sequel we sketch some testing procedures that account for the more general situation of linear dependence over time.

3.2.1 Covariance test

The first robust approach is based on a classical covariance estimator and an estimator of its variance which accounts for serial correlation. Let $p_{\widetilde{Y}} = \mathbb{P}[\widetilde{Y}_t = 1]$ resp. $p_{\widetilde{X}} = \mathbb{P}[\widetilde{X}_t = 1]$ be constant over time, and decompose

$$\widetilde{Y}_t = p_{\widetilde{Y}} + \varepsilon_t^{\widetilde{Y}} \text{ resp. } \widetilde{X}_t = p_{\widetilde{X}} + \varepsilon_t^{\widetilde{X}} \ ,$$

where $\varepsilon_t^{\widetilde{Y}}$ resp. $\varepsilon_t^{\widetilde{X}}$ are binary zero mean random errors which may be serially correlated. Consequently, the null hypothesis that $\text{Cov}(\widetilde{Y}_t, \widetilde{X}_t) = 0$ is equivalent to $\mathbb{E}[\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}] = 0$. Under suitable assumptions (e.g. stationarity and weak dependence of $\{\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}\}_{t=1}^T$) a central limit theorem for $\frac{1}{T}\sum_{t=1}^T \varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}$ holds (Lütkepohl, 2006):

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}} - \mathbb{E}\left[\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}\right]\right) \xrightarrow[T\to\infty]{\mathbb{D}} N\left(0, S\right) \ ,$$

where $S = \sum_{j=-\infty}^{\infty} \text{Cov}\left(\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}, \varepsilon_{t-j}^{\widetilde{Y}} \varepsilon_{t-j}^{\widetilde{X}}\right)$ denotes the approximate asymptotic variance of $\sum_{t=1}^T \varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}/T$. With the consistent estimators $\widehat{p}_{\widetilde{Y}} = \frac{1}{T}\sum_{t=1}^T \widetilde{Y}_t$ and $\widehat{p}_{\widetilde{X}} = \frac{1}{T}\sum_{t=1}^T \widetilde{X}_t$, the unobserved random errors can be estimated consistently by $\widehat{\varepsilon}_t^{\widetilde{Y}} = \widetilde{Y}_t - \widehat{p}_{\widetilde{Y}}$ resp. $\widehat{\varepsilon}_t^{\widetilde{X}} = \widetilde{X}_t - \widehat{p}_{\widetilde{X}}$.

Hence, letting $\overline{\varepsilon^{\widetilde{Y}} \varepsilon^{\widetilde{X}}} = \frac{1}{T}\sum_{t=1}^T \widehat{\varepsilon}_t^{\widetilde{Y}} \widehat{\varepsilon}_t^{\widetilde{X}} = \widehat{\text{Cov}}\left(\widetilde{Y}_t, \widetilde{X}_t\right)$ it can be concluded that

$$CovNW_T = \sqrt{T}\frac{\left(\overline{\varepsilon^{\widetilde{Y}} \varepsilon^{\widetilde{X}}} - \mathbb{E}\left[\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}\right]\right)}{\sqrt{\widehat{S}_T^{NW}}} \xrightarrow[T\to\infty]{\mathbb{D}} N(0,1) \ . \tag{3.2}$$

In (3.2), $\widehat{S}_T^{NW}$ is the heteroscedasticity and autocorrelation consistent variance estimator (Newey and West, 1987) for $\widehat{\text{Cov}}\left(\widetilde{Y}_t, \widetilde{X}_t\right)$

$$\begin{aligned} \widehat{S}_T^{NW} = \widehat{\mathbb{V}}\left[\sqrt{T}\sum_{t=1}^{T} \varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}\right] &= \widehat{\text{Cov}}\left(\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}, \varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}\right) + 2\sum_{q=1}^{Q} \omega(q,Q)\widehat{\text{Cov}}\left(\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}, \varepsilon_{t+q}^{\widetilde{Y}} \varepsilon_{t+q}^{\widetilde{X}}\right) \quad (3.3) \\ \widehat{\text{Cov}}\left(\varepsilon_t^{\widetilde{Y}} \varepsilon_t^{\widetilde{X}}, \varepsilon_{t+q}^{\widetilde{Y}} \varepsilon_{t+q}^{\widetilde{X}}\right) &= \frac{1}{T}\sum_{t=1}^{T-q} \left(\widehat{\varepsilon}_t^{\widetilde{Y}} \widehat{\varepsilon}_t^{\widetilde{X}} - \overline{\varepsilon^{\widetilde{Y}} \varepsilon^{\widetilde{X}}}\right)\left(\widehat{\varepsilon}_{t+q}^{\widetilde{Y}} \widehat{\varepsilon}_{t+q}^{\widetilde{X}} - \overline{\varepsilon^{\widetilde{Y}} \varepsilon^{\widetilde{X}}}\right) \ , \end{aligned}$$

and the weighting function is defined as $\omega(q,Q) = (1 - \frac{q}{Q+1})$. The truncation lag Q can be chosen according to the integer part of $4(T/100)^{2/9}$ (Newey and West, 1994).

Note that under H_0 the squared statistic in (3.2) is equal to the Wald statistic discussed in Holt, Scott and Ewings (1980) or Rao and Scott (1981). The asymptotic covariance matrix of estimated cell proportions is determined by means of the Newey–West approach. We prefer the representation in (3.2) as it allows to test one–sided hypotheses which is particularly useful within the context of directional forecast evaluation.

3.2.2 Static/dynamic regression approach

A test of H_0 which accounts for serial correlation can also be accomplished in the linear regression model. First, consider the static regression model given in (3.1) where the disturbance term ε_t is allowed to be serially correlated. Then, the Newey–West corrected t–statistic for the OLS estimator $\widehat{\beta}_T^{OLS}$ is approximately Gaussian

$$\frac{\widehat{\beta}_T^{OLS} - \beta}{\sqrt{\widehat{\mathbb{V}}_T^{NW}[\widehat{\beta}_T^{OLS}]}} \dot{\approx} N(0,1)$$

(see Breen, Glosten and Jagannathan (1989) for an application).

Another possibility to allow for serial correlation is to dynamically augment model (3.1) with lagged dependent and explanatory variables $\widetilde{X}_t$ resp. $\widetilde{Y}_t$, i.e.:

$$\widetilde{X}_t \;=\; \gamma + \beta\widetilde{Y}_t - \sum_{j=1}^{m}\delta_j\widetilde{Y}_{t-j} + \sum_{j=1}^{m}\rho_j\widetilde{X}_{t-j} + u_t \;. \tag{3.4}$$

Testing H_0 in (3.4) amounts to a test of $\beta = 0$ after correcting for the effects of lagged dependent and explanatory variables. The number of lags m can be chosen according to some information criterion such as the Akaike Information Criterion (AIC). To account for remaining residual autocorrelation the Newey–West corrected t–statistic for $\widehat{\beta}_T^{OLS}$ can be computed (as in PT08). It is again approximately Gaussian. The truncation lag Q is chosen according to the integer part of $4(T/100)^{2/9}$. The tests based on (3.1) and (3.4) are called StatNW resp. DynNW.

3.2.3 Pesaran and Timmerman (2008) test

PT08 propose a more general approach for multicategory variables. Reinterpreting (3.4) as a reduced rank regression, they propose test statistics based on canonical correlations. For the time points $t = 1, ..., T$ and m initial values for $\widetilde{X}_t$ resp. $\widetilde{Y}_t$ model (3.4) can be rewritten as

$$\widetilde{X} = \widetilde{Y}\beta + WB + U \;,$$

with

$$\underset{T\times 1}{\widetilde{X}} = \begin{pmatrix} \widetilde{X}_1 \\ \vdots \\ \widetilde{X}_T \end{pmatrix} \;,\; \underset{T\times 1}{\widetilde{Y}} = \begin{pmatrix} \widetilde{Y}_1 \\ \vdots \\ \widetilde{Y}_T \end{pmatrix} \;,\; \underset{T\times 1}{U} = \begin{pmatrix} u_1 \\ \vdots \\ u_T \end{pmatrix} \;,\; \underset{(2m+1)\times 1}{B} = (\gamma, \delta_1, ..., \delta_m, \rho_1, ..., \rho_m)' \;,$$

$$\underset{T\times(2m+1)}{W} = \begin{pmatrix} 1 & \widetilde{Y}_0 & \widetilde{Y}_{-1} & ... & \widetilde{Y}_{-m+1} & \widetilde{X}_0 & \widetilde{X}_{-1} & ... & \widetilde{X}_{-m+1} \\ \vdots & & & & & & & & \vdots \\ 1 & \widetilde{Y}_{T-1} & \widetilde{Y}_{T-2} & ... & \widetilde{Y}_{T-m} & \widetilde{X}_{T-1} & \widetilde{X}_{T-2} & ... & \widetilde{X}_{T-m} \end{pmatrix} .$$

PT08 show that under the null hypothesis

$$(T-2)S \dot{\approx} \chi^2_{(1)} ,$$

where

$$S = S_{XX}^{-1} S'_{YX} S_{YY}^{-1} S_{YX}, \;\; S_{YY} = \tfrac{1}{T}\widetilde{Y}'M\widetilde{Y}, \;\; S_{YX} = \tfrac{1}{T}\widetilde{Y}'M\widetilde{X}$$
$$S_{XX} = \tfrac{1}{T}\widetilde{X}'M\widetilde{X}, \;\; M = I_T - W\left(W'W\right)^{-1}W', \;\; \widetilde{Y} = (\widetilde{Y}_1, ..., \widetilde{Y}_T)' .$$

In the binary case S is a scalar random variable. Generally, S is a $(c_x - 1) \times (c_x - 1)$–matrix, with c_x being the number of $\widetilde{X}_t$–categories. For finite samples PT08 simulate critical values under H_0 using multinomial sampling. They consider a static and a dynamic version in full analogy to the regression based testing outlined before.

3.2.4 Bootstrap approach

We implement the bootstrap procedure for the covariance test in Section 3.2.1 as it allows both one– and two–sided alternative hypotheses. Moreover, the adaptation to general $r \times c$ contingency tables is possible. The block bootstrap is nowadays commonly accepted as an appropriate bootstrap method if an analyst wants to avoid to impose parametric restrictions on the structure of the data generating process. Künsch (1989), Lahiri (1991), Liu and Singh (1992), Politis and Romano (1992) were among the first to consider the bootstrap for time series. They show that the block bootstrap for time series is a suitable tool to obtain asymptotically valid procedures to approximate distributions of a large class of statistics and weakly dependent data generating processes. Radulovic (1996) proves that consistency of the block bootstrap for the mean usually holds when the statistic is asymptotically normal for a strongly mixing stationary sequence. Götze and Künsch (1996) and Lahiri (1996) cover the asymptotic refinements over the classical normal approximation of the error in rejection probability (ERP) of one–sided tests. Results for two–sided tests are given by Hall and Horowitz (1996), Andrews (2002) and Inoue and Shintani (2006). They demonstrate that the

block bootstrap is more accurate than the normal approximation in terms of ERP for two–sided tests if properly implemented. Various blocking procedures have been proposed. The non–overlapping (NBB) resp. overlapping moving block bootstrap (MBB) were considered by Hall (1985), Carlstein (1986) and Kuensch (1989). Politis and Romano (1992, 1994) introduced the circular block bootstrap (CBB) and the stationary bootstrap (SB). Lahiri (1999) concludes that for estimating the distribution of a studentized statistic the MBB and CBB procedures are more efficient than NBB and SB versions in terms of MSE. The bootstrap sample mean has an expectation equal to the sample mean of the observed series under the CBB which is not the case for the MBB scheme. Hence, for the CBB centering the bootstrap distribution to establish a zero mean distribution is accomplished in the usual way.

To perform a two–sided test of $H_0 : \mathrm{Cov}\left(\widetilde{Y}_t, \widetilde{X}_t\right) = 0$ we investigate two bootstrap approaches for the studentized statistic

$$ST_T \;=\; \frac{\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}} - \mathbb{E}\left[\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}\right]\right)}{\sqrt{\widehat{\mathbb{V}}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}\right]}} \;,$$

where $\widehat{\mathbb{V}}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}\right]$, as given in (3.3), is a consistent estimator of the long run variance of the sample mean of $\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}$. Below we point out that care has to be taken with respect to the choice of the weighting function and the truncation lag. Note that for ease of exposition we do not distinguish between $\varepsilon_t^{\widetilde{Y}}, \varepsilon_t^{\widetilde{X}}$ and $\widehat{\varepsilon}_t^{\widetilde{Y}}, \widehat{\varepsilon}_t^{\widetilde{X}}$. First, consider the observed series $\{\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}\}_{t=1}^{T}$. The circular block bootstrap (CBB) exploits T overlapping blocks of length B given by

$$B_t^{\widetilde{Y}\widetilde{X}} = (\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}, ..., \varepsilon_{t+B-1}^{\widetilde{Y}}\varepsilon_{t+B-1}^{\widetilde{X}}), \;\; t = 1, ..., T \;.$$

Observations $\varepsilon_t^{\widetilde{Y}}\varepsilon_t^{\widetilde{X}}$ for $r > T$ are wrapped around in a circle, i.e. $\varepsilon_{T+b}^{\widetilde{Y}}\varepsilon_{T+b}^{\widetilde{X}} = \varepsilon_b^{\widetilde{Y}}\varepsilon_b^{\widetilde{X}}$ for $1 \leq b \leq B$. Let the integer part of T/B, $[T/B]$, be the number of blocks K which are drawn randomly with replacement from the set of blocks $B_t^{\widetilde{Y}\widetilde{X}}$. Each of the drawn blocks, $k = 1, ..., K$, is denoted by $\xi_k^{\widetilde{Y}\widetilde{X}} = (\xi_{k,1}^{\widetilde{Y}\widetilde{X}}, ..., \xi_{k,B}^{\widetilde{Y}\widetilde{X}})$. Concatenating all $\xi_{k,b}^{\widetilde{Y}\widetilde{X}}$ in a vector defines the bootstrap sample $V_1^*, ..., V_L^*$. Thus the length of the bootstrap sample is $L = KB \leq T$, and the bootstrap sample average is

$$\bar{V}_L^* = \frac{1}{L}\sum_{t=1}^{L} V_t^* = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{b}\sum_{b=1}^{B}\xi_{k,b}^{\widetilde{Y}\widetilde{X}}\right) \;.$$

Under CBB sampling (which implies a measure P^*_{CBB1}) it can be shown that

$$\mathbb{E}^*_{CBB1}\left[\bar{V}^*_L\right] = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}} \ ,$$

where $\mathbb{E}^*_{CBB1}$ is the expectation under P^*_{CBB1}. Davison and Hall (1993) demonstrate that the block bootstrap for studentized statistics provides an improvement in asymptotic accuracy when applied properly. In particular, the naive studentization based on plugging in the bootstrapped sample into the formula for the long run variance estimator $\widehat{\mathbb{V}}[\bullet]$, i.e.

$$\frac{\sqrt{L}\left(\bar{V}^*_L - \mathbb{E}^*_{CBB1}\left[\bar{V}^*_L\right]\right)}{\sqrt{\widehat{\mathbb{V}}\left[\sqrt{L}\bar{V}^*_L\right]}} \ ,$$

yields a bootstrap approximation which maybe less accurate than the classical normal approximation. For ERPs of one–sided tests asymptotic refinements are obtained when studentization is accomplished by means of the variance of the rescaled bootstrap average under P^*_{CBB1} (Lahiri, 1991 and 1996, Götze and Künsch, 1996). This is given by

$$\mathbb{V}^*_{CBB1}\left[\sqrt{L}\bar{V}^*_L\right] = \frac{B}{T}\sum_{t=1}^{T}\left[\left(\frac{1}{B}\sum_{b=1}^{B}\varepsilon^{\tilde{Y}}_{b+t-1}\varepsilon^{\tilde{X}}_{b+t-1}\right) - \mathbb{E}^*_{CBB1}\left[\bar{V}^*_L\right]\right]^2 \ .$$

For two–sided tests the studentization by means of $\mathbb{V}^*_{CBB1}\left[\sqrt{L}\bar{V}^*_L\right]$ does not yield a superior performance over the normal approximation. Lahiri (1992) and Hall and Horowitz (1996) introduce correction factors to obtain refinements for both one– and two–sided symmetric tests. In particular, they define the bootstrap statistic as

$$ST^*_{T,CBB1} = \frac{\sqrt{L}\left(\bar{V}^*_L - \mathbb{E}^*_{CBB1}\left[\bar{V}^*_L\right]\right)}{\sqrt{\widehat{\mathbb{V}}\left[\sqrt{L}\bar{V}^*_L\right]}}\sqrt{\frac{\mathbb{V}_{CBB1}\left[\frac{1}{\sqrt{L}}\sum_{t=1}^{L}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}}\right]}{\mathbb{V}^*_{CBB1}\left[\sqrt{L}\bar{V}^*_L\right]}} \ , \tag{3.5}$$

where $\mathbb{V}_{CBB1}\left[\frac{1}{\sqrt{L}}\sum_{t=1}^{L}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}}\right]$ is the bootstrap analog of $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}}\right]$. Hence, the former is given by (3.3) with a weighting function $\omega(q,Q)=1$ and truncation lag $Q=T-1$.

Next, a bootstrap procedure explicitly accounting for the independence of $\varepsilon_t^{\tilde{Y}}$ and $\varepsilon_t^{\tilde{X}}$ under the null hypothesis is outlined. We randomly resample with replacement K circular blocks of $\varepsilon_t^{\tilde{X}}$

$$B_t^{\tilde{X}} = (\varepsilon_t^{\tilde{X}}, ..., \varepsilon_{t+B-1}^{\tilde{X}}), \ \ t = 1, ..., L \ .$$

Concatenating the resampled blocks $\xi_k^{\tilde{X}} = (\xi_{k,1}^{\tilde{X}}, ..., \xi_{k,B}^{\tilde{X}})$ in a vector, the bootstrap sample average is given by

$$\bar{V}_L^* = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{B}\sum_{b=1}^{B}\varepsilon_{B(k-1)+b}^{\tilde{Y}}\xi_{k,b}^{\tilde{X}}\right) .$$

This resampling approach implies

$$\begin{aligned}
\mathbb{E}_{CBB2}^*\left[\bar{V}_L^*\right] &= \overline{\varepsilon_L^{\tilde{Y}}}\ \overline{\varepsilon_L^{\tilde{X}}} \\
\mathbb{V}_{CBB2}^*\left[\sqrt{L}\bar{V}_L^*\right] &= \frac{L}{K^2}\sum_{k=1}^{K}\left(\frac{1}{L}\sum_{t=1}^{L}\left[\left(\frac{1}{B}\sum_{b=1}^{B}\varepsilon_{B(k-1)+b}^{\tilde{Y}}\varepsilon_{b+t-1}^{\tilde{X}}\right) - \overline{\varepsilon_L^{\tilde{X}}}\,S_k^{\tilde{Y}}\right]^2\right) ,
\end{aligned}$$

where $\overline{\varepsilon_L^{\tilde{Y}}} = (1/L)\sum_{t=1}^{L}\varepsilon_t^{\tilde{Y}}$, $\overline{\varepsilon_L^{\tilde{X}}} = (1/L)\sum_{t=1}^{L}\varepsilon_t^{\tilde{X}}$ and $S_k^{\tilde{Y}} = (1/B)\sum_{b=1}^{B}\varepsilon_{B(k-1)+b}^{\tilde{Y}}$. Accordingly, the bootstrap statistic is

$$ST_{T,CBB2}^* = \frac{\sqrt{L}\left(\bar{V}_L^* - \mathbb{E}_{CBB2}^*\left[\bar{V}_L^*\right]\right)}{\sqrt{\widehat{\mathbb{V}}\left[\sqrt{L}\bar{V}_L^*\right]}}\sqrt{\frac{\mathbb{V}_{CBB2}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}}\right]}{\mathbb{V}_{CBB2}^*\left[\sqrt{L}\bar{V}_L^*\right]}} . \tag{3.6}$$

Note that for this bootstrap scheme the bootstrap analog to $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}}\right]$ is given by

$$\begin{aligned}
\mathbb{V}_{CBB2}\left[\frac{1}{\sqrt{L}}\sum_{t=1}^{L}\varepsilon_t^{\tilde{Y}}\varepsilon_t^{\tilde{X}}\right] &= \widehat{\mathrm{Cov}}\left(\varepsilon_t^{\tilde{Y}},\varepsilon_t^{\tilde{Y}}\right)\widehat{\mathrm{Cov}}\left(\varepsilon_t^{\tilde{X}},\varepsilon_t^{\tilde{X}}\right) \\
&\quad +2\sum_{q=1}^{T-1}\widehat{\mathrm{Cov}}\left(\varepsilon_t^{\tilde{Y}},\varepsilon_{t+q}^{\tilde{Y}}\right)\widehat{\mathrm{Cov}}\left(\varepsilon_t^{\tilde{X}},\varepsilon_{t+q}^{\tilde{X}}\right) \\
&\quad -L\left(\overline{\varepsilon_L^{\tilde{Y}}}\ \overline{\varepsilon_L^{\tilde{X}}}\right)^2 ,
\end{aligned}$$

which explicitly accounts for the independence of $\varepsilon_t^{\tilde{Y}}$ and $\varepsilon_t^{\tilde{X}}$.

3.2.5 Implementation of the bootstrap approach

The choice of the kernel function $\omega(q, Q)$ is crucial for the bootstrap to provide better approximations than the classical normal approximation. For one–sided tests, Götze and Künsch (1996) show that for all kernels but the Bartlett kernel improvements in ERP's can be obtained when $B = Q = O(T^{1/4})$. Moreover, they point out that their results also hold for other choices of $B \leq Q$. Hall and Horowitz (1996) and Andrews (2002) consider approximation errors for two–sided symmetric tests when the truncated kernel is used. Inoue and Shintani (2006) extend these results to show that for kernels such as the truncated, trapezoidal or Parzen kernel the bootstrap yields refinements for two–sided

symmetric tests when $B = Q = O(T^{1/3})$. They also point out that their results hold if block sizes B are proportional to the truncation parameter Q. Otherwise the rate of the bootstrap approximation error is determined by the faster rate of B and Q. Hence, in our analysis we choose the truncated kernel and set $B = Q$. The choice of the truncated kernel does not guarantee that the variance estimator is positive. Yet, for positively persistent time series this problem is not as crucial as compared to data exhibiting negative serial correlation.

In order to implement a block bootstrap the block length parameter B has to be specified. Various approaches to determine optimal block sizes have been proposed. Hall, Horowitz and Jing (1995) derive optimal block sizes based on an asymptotic mean squared error criterion for bias/variance estimation or one– and two–sided distribution estimation. They show that optimal block lengths are $O(T^{1/3}), O(T^{1/4})$ and $O(T^{1/5})$, respectively. Zvingelis (2001) determines an asymptotically optimal block length minimizing the asymptotic ERP of one– and two–sided tests. He concludes that the optimal block sizes are $O(T^{1/4})$ and $O(T^{1/3})$, respectively. The constants of proportionality depend on, e.g., the autocovariance function of the DGP. Politis and White (2004) derive for the CBB scheme an explicit expression of the optimal block length B_{opt} for an AR(1)–process when interest focuses on bias/variance or distribution function estimation. They show that the optimal block size increases with the autocorrelation coefficient.

Relying on the result of Zvingelis (2001) an adaptation of the data based block length selection procedure of Hall, Horowitz and Jing (1995) targeting the empirical ERP criterion is straightforward. In particular, we compute the empirical ERP of the bootstrap test for all subsamples of length $\widetilde{T} < T$ and a grid of selected block lengths. Given the block length $B_{\widetilde{T}}$, for which the empirical ERP is closest to the nominal significance level, the estimated optimal block length for a sample of size T is then obtained from $\widehat{B}_{\text{opt}} = (T/\widetilde{T})^{1/3} B_{\widetilde{T}}$ for a two–sided test.

4 Simulation results

In order to shed light on the small sample properties of the test procedures presented above, we carry out a simulation study. We document the MC design and describe the size and size–adjusted power results, in turn.

4.1 Design

To simulate Bernoulli serially correlated random variables, we consider the stationary 2–dimensional VAR(1) process

$$\begin{pmatrix} Z_{1t} \\ Z_{2t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & 0 \\ 0 & \phi_{22} \end{pmatrix} \begin{pmatrix} Z_{1t-1} \\ Z_{2t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} ,$$

with $|\phi_{ii}| < 1,\ i = 1, 2$, and

$$\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim NID \left[0, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right] .$$

Defining $\sigma_i^2 = (1 - \phi_{ii}^2),\ i = 1, 2$, $\sigma_{12} = \rho(1 - \phi_{11}\phi_{22})$ and $\phi_{11} = \phi_{22}$, the univariate processes, Z_{1t} and Z_{2t}, have unit variance $\mathbb{V}[Z_{it}] = 1,\ i = 1, 2$ and serial correlation $\text{Corr}(Z_{it}, Z_{it-j}) = \phi_{ii}^j, i = 1, 2$. Moreover, the contemporaneous cross covariance/correlation is given by $\text{Cov}(Z_{1t}, Z_{2t}) = \text{Corr}(Z_{1t}, Z_{2t}) = \rho$. Hence, cross sectional dependence and serial correlation are specified by selection of $|\rho| < 1$ and ϕ_{11}, respectively. Finally, let

$$\widetilde{X}_t = 1\left(Z_{1t} > 0\right) \text{ and } \widetilde{Y}_t = 1\left(Z_{2t} > 0\right) .$$

For cross sectional independence ($\rho = 0.0$) and medium and strong cross sectional dependence ($\rho = 0.5$ and $\rho = 0.8$) we simulate 5000 Monte Carlo replications of the process with no, medium and strong serial dependence ($\phi_{11} \in \{0.0, 0.5, 0.8\}$). We consider samples of size $T \in \{20, 50, 100, 500, 1000\}$. For each Monte Carlo replication we use 100 initializing presample values. The Fisher test is implemented as described in Henrikkson and Merton (1981). For the dynamic regression approach a maximum lag of 4 is allowed when choosing the lag order by means of the AIC. The truncation lag in the Newey–West estimation procedure is given by the integer part of $4(T/100)^{2/9}$. Finally, for the bootstrap approach we choose B, naively, as the nearest integer to $T^{1/3}$.

In our simulations a naive choice of the block size leads to rejection frequencies smaller than the nominal level of 5% for $T \geq 100$. Thus, for $T \geq 100$ we also choose the block length using the data based selection approach of Hall, Horowitz and Jing (1995). More precisely, for $T = 100$ the subsample length and the grid of block lengths are $\widetilde{T} = 30$ and $B_{\text{grid}} = \{3, 4, ..., 7\}$. For $T = 500, 1000$ we set $\widetilde{T} = 100$ and $B_{\text{grid}} = \{3, 4, ..., 15\}$.

4.2 Rejection frequencies under H_0

First, we describe the results for the case of cross sectional independence ($\rho = 0$) and no serial correlation ($\phi_{11} = 0$). The nominal significance level is 5%. Notably, results for other nominal levels are qualitatively identical. From the upper panel of Table 2 it can be inferred that the classical χ^2, the PT92, the PT08 and the bootstrap test perform very well and have empirical rejection frequencies very close to the nominal 5% level for all sample sizes considered. Fisher's test is seriously oversized in small and medium sample sizes but rejection rates converge to the nominal level as T increases. The small sample size distortion is possibly due to the fact that the simulation design does not guarantee fixed row and column marginals. The CovNW, StatNW and DynNW test procedures are also markedly oversized in small samples but approach empirical rejection frequencies close to 5% for increasing T. Accounting for serial correlation when there is none, does not pay off in small samples. Yet, it is not surprising that correctly assuming serial independence leads to an improved performance.

The medium panel of Table 2 displays empirical rejection frequencies under moderate serial correlation ($\phi_{11} = 0.5$). It reveals some size distortions for all but the bootstrap test. The χ^2 and the PT92 tests share similar rejection frequencies between 7% and 8% for all sample sizes considered. Fisher's test is seriously oversized in small samples with a rejection frequency of $\approx 8.5\%$. Among the three test procedures relying on the Newey–West variance estimator, the CovNW approach uncovers smallest size distortions for small samples. Empirical rejection frequencies of robust tests converge to the nominal level of 5% for all of these tests. Using the PT08 test H_0 is oversized in small samples but for medium and large samples the test has the correct rejection rate.

Introducing strong serial correlation ($\phi_{11} = 0.8$), the lower panel of Table 2 indicates that size distortions are severe for those tests which do not account for serial correlation. The χ^2, Fisher and PT92 tests are massively oversized for all sample sizes considered. Relative rejection frequencies appear to converge to $\approx 20\%$. The rejection frequency of the CovNW, StatNW and DynNW approaches is far too high in small samples but stabilizes $\approx 7\%$. The PT08 test is for small samples oversized (>10%) but as $T \geq 100$ it has appropriate rejection frequency. Among all tests considered the bootstrap approach performs best. It reveals (if any) small size distortions with empirical rejection frequencies close to the nominal level. For

$T \geq 100$ the simulations for the alternative block length selection reveal a robust performance with rejection frequencies around 5%.

In summary, the bootstrap approach turns out to offer a remarkably robust performance. Its implied empirical size is close to the nominal level under serial independence and in the presence of serial correlation for all sample sizes considered. The PT08 approach is robust to serial dependence for medium and large sample sizes but reveals size distortions if $T < 100$.

4.3 Size–adjusted power

Table 2 documents the size–adjusted power results for selected scenarios of serial correlation ($\phi_{11} = 0.0, 0.5, 0.8$) when the cross correlation is $\rho = 0.5$ resp. $\rho = 0.8$. Some general conclusions can be drawn for all tests considered. The power decreases with increasing serial correlation. In the presence of serial dependence concordant observations of $\widetilde{X}_t$ and $\widetilde{Y}_t$ are more likely. Hence, it is more difficult to isolate the effects of cross sectional and serial dependence. Furthermore and most reasonably, size–adjusted power increases with increasing cross correlation.

While for sample sizes larger than 100 the power performance is very similar across the various test procedures, there are some differences for smaller sample sizes. For $T = 20, 50$ the χ^2, Fisher's, the PT92 and the PT08 test are somewhat more powerful than the CovNW, StatNW, DynNW and the bootstrap test. For example, while the former tests have a power close to 80% the latter reject slightly less frequently in less than 75% of the cases when $\rho = 0.8,\ \phi_{11} = 0.8$ and $T = 50$. The power of the bootstrap test is despite its non–parametric nature very appealing. It is close to the power of the remaining tests. For example, when $\rho = 0.8,\ \phi_{11} = 0.8$ and $T = 50$, the rejection rate of the bootstrap approach is 60%.

$\phi_{11} = 0.0$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.055	**0.172**	**0.063**	**0.120**	**0.111**	**0.122**	**0.067**	**0.057**
50	0.054	**0.103**	0.055	**0.079**	**0.075**	**0.075**	0.049	0.055
100	0.052	**0.082**	0.053	**0.063**	**0.061**	**0.061**	0.049	0.050
500	0.056	**0.067**	0.056	**0.058**	**0.057**	**0.057**	0.054	0.049
1000	0.050	**0.059**	0.050	0.054	0.054	0.054	0.051	0.045

$\phi_{11} = 0.5$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	**0.081**	**0.246**	**0.090**	**0.135**	**0.150**	**0.161**	**0.091**	**0.064**
50	**0.085**	**0.148**	**0.088**	**0.098**	**0.094**	**0.093**	**0.074**	0.048
100	**0.081**	**0.117**	**0.082**	**0.072**	**0.071**	**0.067**	**0.057**	0.049
500	**0.086**	**0.102**	**0.086**	**0.063**	**0.063**	0.054	0.053	0.052
1000	**0.087**	**0.097**	**0.088**	**0.062**	**0.061**	0.053	0.053	0.053

$\phi_{11} = 0.8$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	**0.155**	**0.507**	**0.163**	**0.161**	**0.263**	**0.226**	**0.141**	**0.060**
50	**0.198**	**0.301**	**0.204**	**0.152**	**0.168**	**0.099**	**0.092**	**0.043**
100	**0.219**	**0.276**	**0.220**	**0.129**	**0.131**	**0.068**	0.056	0.051
500	**0.239**	**0.262**	**0.240**	**0.097**	**0.097**	0.051	0.049	0.053
1000	**0.242**	**0.257**	**0.242**	**0.094**	**0.094**	0.051	0.053	0.045

Table 1. Empirical rejection frequencies under H_0 ($\rho = 0.0$) and nominal significance level 5%. Different serial correlation parameters $\phi_{11} \in \{0.0, 0.5, 0.8\}$ and sample sizes $T \in \{20, 50, 100, 500, 1000\}$ are considered. χ^2 and FE denote the χ^2– and Fisher's exact test, respectively. Moreover, CovNW, StatNW, DynNW denote the covariance test and the tests based on the static and dynamic regression approaches using the Newey–West variance estimator. Corresponding naively chosen block sizes are $3, 4, 5, 8, 10$ when $\rho = 0.0$. When $\rho = 0.5, 0.8$ block sizes are 3 and 4 for $T = 20$ and $T = 50$. For $T \geq 100$ block sizes are determined by means of the approach proposed by Hall, Horowitz and Jing (1995). Bold figures are not within the 95% confidence interval given by $[\alpha \pm 2\sqrt{\alpha(1-\alpha)/5000}]$, where $\alpha = 0.05$.

$\rho = 0.5, \phi_{11} = 0.0$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.296	0.268	0.296	0.265	0.259	0.239	0.270	0.167
50	0.681	0.675	0.681	0.638	0.637	0.633	0.674	0.439
100	0.931	0.931	0.931	0.925	0.926	0.925	0.931	0.844
500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\rho = 0.8, \phi_{11} = 0.0$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.775	0.742	0.775	0.703	0.699	0.663	0.740	0.443
50	0.995	0.995	0.995	0.989	0.990	0.990	0.995	0.931
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\rho = 0.5, \phi_{11} = 0.5$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.233	-	0.236	0.192	0.177	0.176	0.192	0.134
50	0.558	0.552	0.560	0.523	0.520	0.504	0.535	0.400
100	0.857	0.858	0.857	0.848	0.851	0.835	0.853	0.708
500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\rho = 0.8, \phi_{11} = 0.5$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.671	-	0.673	0.537	0.540	0.524	0.604	0.317
50	0.978	0.977	0.978	0.968	0.969	0.963	0.975	0.891
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\rho = 0.5, \phi_{11} = 0.8$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.158	-	0.155	0.127	-	-	0.144	0.081
50	0.303	-	0.316	0.282	0.282	0.247	0.301	0.233
100	0.570	0.577	0.572	0.545	0.550	0.519	0.568	0.429
500	0.996	0.998	0.996	0.997	0.997	0.998	0.999	0.991
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\rho = 0.8, \phi_{11} = 0.8$

T	χ^2	FE	PT 92	Cov NW	Stat NW	Dyn NW	PT 08	CBB2
20	0.453	-	0.449	0.339	-	-	0.412	0.160
50	0.818	-	0.827	0.734	0.766	0.699	0.796	0.628
100	0.979	0.980	0.980	0.970	0.976	0.968	0.981	0.909
500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2. Size–adjusted power. Different cross sectional correlation parameters $\rho \in \{0.5, 0.8\}$, serial correlation parameters $\phi_{11} \in \{0.0, 0.5, 0.8\}$ and sample sizes $T \in \{20, 50, 100, 500, 1000\}$ are considered. Note, no size–adjusted power is reported for Fisher's, the StatNW and the DynNW test in some cases. Due to the discreteness of the data it happens that at a nominal significance level of 0.1% the empirical size is 8% or larger. For further notes see Table .

5 Empirical applications

To illustrate the application of the test procedures and highlight the importance of accounting for serial correlation in applied work, we consider two empirical examples.

5.1 A large sample case

We apply the χ^2, Fisher's, the PT92, the PT08 and the bootstrap test to analyze directional forecasts for selected EURIBOR swap rates. Blaskowitz and Herwartz (2008) consider $h = 1, 5, 10, 15$ days ahead ex–ante forecast for the EURIBOR swap term structure. Based on a battery of factor models they adaptively combine models to produce 1778 daily forecasts for the 2yr swap rate from April 19th, 2000, till mid February / beginning of March 2007 (depending on the forecast horizon h). We consider forecasts obtained from the most preferable Median strategy.

For comparison purposes some benchmark models are also considered. Namely, an AR(1) model and a variant of the term structure model proposed by Diebold and Li (2006) are fitted by means of rolling windows of 42 daily observations (see Blaskowitz and Herwartz, 2008) for details. The benchmark strategies are denoted by AR resp. DL.

Table 3 illustrates the extent of serial correlation present in realized and forecasted directions (up–/downward movements) of the 2yr EURIBOR swap rate. Apart from the realized directions of the 2yr swap rate for one day ahead forecasts, all remaining series are highly and significantly serially correlated. Moreover, the higher the horizon, the stronger the serial dependence. For forecast horizons $h = 5, 10, 15$ first order correlations for outcomes in directions are high, about 0.6, 0.7, 0.8, respectively. Correlations decrease to less than 0.1 at lag 20. For forecasted directions, first order correlations are between 0.75 and 0.93 for $h = 5, 10, 15$ and remain high (above ≈ 0.4) at all lags considered. This evidence suggests that commonly applied procedures to test for the value of directional forecasts in the sense of Merton (1981) are inadequate for all but the one day ahead forecasts of the 2yr swap rate.

To analyze the value of EURIBOR swap rate forecasts, Table 4 shows empirical estimates of covariances and HM statistics for various forecast exercises. It can be seen that the forecasts of all models have positive value. Moreover, Table 4 provides the results for testing $H_0 : \mathrm{Cov}(\widetilde{Y}_t, \widetilde{X}_t) = 0$ against $H_1 : \mathrm{Cov}(\widetilde{Y}_t, \widetilde{X}_t) \neq 0$ for various significance levels $\alpha \leq 0.20$. Using traditional test procedures the null of no value is rejected at a 1% significance level for

all forecast exercises except for $h = 1$ AR and DL forecasts. For the latter, H_0 is rejected at nominal levels between 11% and 15%. Conclusions drawn from the serial correlation robust test procedures are different in some cases. The discrepancy becomes more apparent the larger the serial correlation. Test decisions for $h = 1$ generally agree for all procedures. Yet, striking differences in significance are obtained for the 5 day ahead forecasts for the DL model as well as for the 10 day ahead forecasts for the Median strategy and the DL model. Note, for the bootstrap test we used a the data based block length selection method of Hall, Horowitz and Jing (1995) as described in Section 3.2.4. Using a naive block choice $B = [1778^{1/3}] = 12$ does not change the conclusions.

	$h=1$	$h=5$	$h=10$	$h=15$	$h=1$	$h=5$	$h=10$	$h=15$
	Realized Directions				MedStrat			
1	-0.040	**0.605**	**0.749**	**0.802**	**0.365**	**0.782**	**0.896**	**0.931**
5	0.015	**0.070**	**0.393**	**0.524**	**0.323**	**0.741**	**0.841**	**0.868**
10	-0.009	**0.056**	**0.063**	**0.272**	**0.270**	**0.699**	**0.805**	**0.828**
15	0.015	0.044	**0.050**	**0.086**	**0.239**	**0.657**	**0.754**	**0.775**
20	-0.012	**0.060**	**0.083**	**0.076**	**0.208**	**0.609**	**0.700**	**0.721**
	AR				DL			
1	**0.546**	**0.815**	**0.851**	**0.864**	**0.097**	**0.676**	**0.832**	**0.863**
5	**0.503**	**0.693**	**0.714**	**0.720**	**0.097**	**0.529**	**0.710**	**0.726**
10	**0.410**	**0.603**	**0.616**	**0.625**	**0.098**	**0.409**	**0.596**	**0.629**
15	**0.407**	**0.560**	**0.564**	**0.573**	0.010	**0.272**	**0.523**	**0.546**
20	**0.368**	**0.489**	**0.499**	**0.501**	**0.068**	**0.211**	**0.418**	**0.458**

Table 3. Serial correlations of realized and forecasted directions of EURIBOR swap rates. Bold numbers are significant at a 5% significance level. Critical values are $\pm 2/\sqrt{1778} \approx \pm 0.047$.

	MedStrat				AR				DL			
	$h=1$	$h=5$	$h=10$	$h=15$	$h=1$	$h=5$	$h=10$	$h=15$	$h=1$	$h=5$	$h=10$	$h=15$
Cov	0.023	0.035	0.036	0.042	0.009	0.029	0.038	0.045	0.002	0.013	0.029	0.037
HM	1.091	1.139	1.143	1.170	1.036	1.118	1.152	1.182	1.010	1.052	1.115	1.148
χ^2	1%	1%	1%	1%	14%	1%	1%	1%	15%	1%	1%	1%
FE	1%	1%	1%	1%	13%	1%	1%	1%	11%	1%	1%	1%
PT92	1%	1%	1%	1%	14%	1%	1%	1%	15%	1%	1%	1%
PT08	1%	2%	NR	1%	14%	1%	5%	13%	15%	NR	NR	2%
CBB2	1%	1%	6%	6%	NR	3%	5%	4%	NR	10%	11%	7%

Table 4. Covariances, HM statistics and test results for various significance levels $\alpha \leq 0.2$ are provided. NR indicates that H_0 cannot be rejected at the 20% significance level.

5.2 A small sample case

Moreover, we investigate the stock return predictions analyzed in Herwartz and Morales (2008). Based on a panel asset pricing model they determine $h = 3, 6$ month ahead forecast of returns of Germany's DAX30, Italy's MIB30 and Norway's OBX25. We focus on the most recent 50 forecasts which cover the period 06/2000 to 01/2005 (depending on the forecast horizon). Positive/negative realized resp. forecasted returns are considered as up–/downward movements.

As can be seen from Table 5 the covariance and HM statistic for the 6 months ahead forecasts of Norway's OBX25 returns are quite large, around 0.15 resp. 1.6. Even if serial correlations are significant at least up to lags 4 and 6 for realized and forecasted directions all test procedures reject at low significance levels. As the test statistic is high, any test should reject the null and the impact of serial correlation should be negligible.

For Germany's DAX30 both the 3 and 6 month ahead forecasts have a rather low value. Covariances and HM statistics are about 0.05 resp. 1.2. Moreover, there is no marked serial correlation beyond lag 4. Thus, similar decisions are inferred from all tests. The null hypothesis is not rejected at conventional significance levels.

The 3 and 6 month ahead forecasts of the MIB30 and the 3 month ahead forecasts of the OBX25 have a moderate value, with covariances between 0.08 and 0.09 and HM statistics between 1.3 and 1.4. Serial correlations are significant up to lags 4 and 6. In such a situation accounting for serial correlation is important when testing for the value of directional forecasts. While all the classical tests reject the null hypothesis, the serial

correlation robust procedures yield a downgrading of the forecast's economic value. The bootstrap test is carried out using $B = 4$. Alternative choices of $B = 2$ and $B = 6$ provide the same results.

Serial correlations

$h = 3$							$h = 6$					
realized directions			forecasted directions				realized directions			forecasted directions		
Ger	Ita	Nor	Ger	Ita	Nor	lag	Ger	Ital	Nor	Ger	Ita	Nor
0.533	**0.619**	**0.497**	**0.394**	**0.628**	**0.661**	1	**0.760**	**0.650**	**0.661**	0.071	**0.694**	**0.740**
0.151	0.199	**0.314**	0.251	**0.550**	**0.529**	4	0.236	**0.466**	**0.325**	-0.019	**0.458**	**0.560**
0.039	0.117	0.075	0.259	**0.453**	**0.371**	6	0.214	**0.389**	0.204	0.113	0.261	**0.360**
0.063	0.135	0.053	0.065	0.258	0.270	9	0.013	0.206	-0.015	0.022	0.143	0.060
-0.202	-0.045	-0.049	0.125	0.063	0.089	12	-0.187	0.104	0.045	-0.131	-0.095	-0.040

Test statistics and test results

	$h = 3$				$h = 6$		
	Ger	Ita	Nor		Ger	Ita	Nor
Cov	0.047	0.082	0.078	Cov	0.055	0.090	0.150
HM	1.200	1.350	1.312	HM	1.254	1.403	1.600
χ^2	17%	2%	3%	χ^2	9%	1%	1%
FE	10%	1%	2%	FE	4%	1%	1%
PT92	17%	2%	3%	PT92	8%	1%	1%
PT08	NR	NR	NR	PT08	9%	NR	1%
CBB2	NR	9%	NR	CBB2	12%	16%	4%

Table 5. Upper panel shows serial correlations of realized and forecasted directions of European stock market returns. Bold numbers are significant at a 5% significance level. Critical values are $\pm 2/\sqrt{50} \approx \pm 0.283$. The lower panel provides covariances, HM statistics and test results for significance levels $\alpha \leq 0.2$. NR indicates that H_0 cannot be rejected at the 20% significance level.

6 Conclusions

Commonly applied procedures to test for the value of directional forecasts suffer from marked size distortions in the presence of serial correlation. As this issue is highly relevant for economic applications, we summarized existing procedures and proposed a simple statistic for which we implement a bootstrap approach. By means of a Monte Carlo simulation we find that the bootstrap test reveals only minor size distortions in small samples as opposed to traditional procedures and retains appealing power. For medium and large sample sizes, the dynamically augmented maximum correlation test proposed in Pesaran and Timmerman (2008) represents an alternative approach with correct size and promising power. In two empirical examples we illustrate the relevance and application of serial correlation robust test procedures for small as well as for large sample sizes.

A particular merit of the investigated test statistic is that it allows for both one–sided and two–sided alternative hypotheses. Moreover, since its square is equal to a Wald statistic under the null hypothesis the test procedure can be easily extended to general $r \times c$ contingency tables. In this framework, the generalized test of market timing as proposed in Pesaran and Timmermann (1992) can be dealt with readily. In principle, the remaining test procedures summarized in this paper can be subjected to resampling. Yet, for the reasons outlined above we focuse on the covariance test statistic and leave it for further research to develop bootstrap algorithms for the other tests.

An appropriate choice of the block length is important for a proper bootstrap test. Our simulations reveal that a naive choice based on the fact that the optimal block length is $O(T^{1/3})$ results in a slightly undersized bootstrap scheme for large sample sizes. Adapting the data based block length selection procedure of Hall, Horowitz and Jing (1995) yields empirical rejection frequencies close to the nominal size. Additional improvements can be expected by a block size selection procedure that accounts for size and power considerations. We regard the latter issue to merit further reflection.

References

Agresti, A., 'A Survey of Exact Inference for Contingency Tables', *Statistical Science*, **7**, 131–153 (1992).

Altham, S., 'Serial Dependence in Contingency Tables', *Journal of the Royal Statistical Society* **Ser. B, 45**, 100–106 (1983).

Anatolyev, S., 'A Unifying View of Some Nonparametric Predictability Tests' *Working Paper, New Economic School, Moscow, Russia* (2006).

Anatolyev, S. and A. Gerko, 'A Trading Approach to Testing for Predictability' *Journal of Business and Economic Statistics*, **23**, 455–461 (2005).

Andrews, D.W.K., 'Higher–Order Improvements of a Computationally Attractive k–Step Bootstrap for Extremum Estimators' *Econometrica*, **70**, 119–162 (2002).

Artis, M.J., 'How Accurate are the IMF's Short–Term Forecasts? Another Examination of the World Economic Outlook', *International Monetary Fund, Working Paper No. 96/89* (1996).

Ash, J.C.K., D.J. Smith and S. M. Heravi, 'Are OECD Forecasts Rational and Useful?: A Directional Analysis', *International Journal of Forecasting*, **14**, 381–391 (1998).

Ashiya, M., 'The Directional Accuracy of 15–Months–Ahead Forecasts Made by the IMF', *Applied Economics Letters*, **10**, 331–333 (2003).

Ashiya, M., 'Are 16–Month–Ahead Forecasts Useful?: A Directional Analysis of Japanese GDP Forecasts', *Journal of Forecasting*, **25**, 201–207 (2006).

Bartlett, M.S., 'The Frequency Goodness of Fit Test for Probability Chains', *Proceedings of the Cambridge Philosophical Society*, **47**, 86–95 (1951).

Blaskowitz, O. and H. Herwartz, 'Adaptive Forecasting of the EURIBOR Swap Term Structure', forthcoming in *Journal of Forecasting* (2008).

Breen, W., L.G. Glosten and R. Jagannathan, 'Economic Significance of Predictable Variations in Stock Index Returns', *Journal of Finance*, **44**, 1177–1189 (1989).

Carlstein, E., 'The Use of Subseries Methods for Estimating the Variance of a General Statistic from Stationary Data', *Annals of Statistics*, **14**, 1171–1179 (1982).

Cicarelli, J., 'A New Method for Evaluating the Accuracy of Economic Forecasts', *Journal of Macroeconomics*, **4**, 469–475 (1982).

Cox, D.R. and D.V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London (1974).

Cumby, R.E. and D.M. Modest, 'Testing For Market Timing Ability', *Journal of Financial Economics*, **19**, 169–189 (1987).

Davison A.C. and P. Hall, 'On Studentizing and Blocking Methods for Implementing the Bootstrap with Dependent Data', *Australian Journal of Statistics*, **35**, 215–224, (1993).

Diebold, F.X., *Elements of Forecasting*, Cincinnati, South-Western College Publishing (2007).

Diebold, F.X. and R. Mariano, 'Comparing Predictive Accuracy', *Journal of Business & Economic Statistics*, **13**, 253–263, (1995).

Easaw, J.Z., D. Garratt and S.M. Heravi, 'Does Consumer Sentiment Accurately Forecast UK Household Consumption? Are There Any Comparisons to be Made With the US?', *Journal of Macroeconomics*, **27**, 517–532 (2005).

Fisher, R.A., *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh (1934).

Gençay, R., 'Optimization of Technical Trading Strategies and the Profitability in Security Markets', *Economics Letters*, **59**, 249–254 (1998).

Götze F. and H. Künsch, 'Second–Order Correctnes of the Blockwise Bootstrap for Stationary Observations', *Annals of Statistics*, **24**, 1914–1933 (1996).

Gradojevic, N. and J. Yang, 'Non–linear, Non–fundamental Exchange Rate Forecasting', *Journal of Forecasting*, **25**, 227–245 (2006).

Granger, C.W.J. and M.H. Pesaran, 'Economic and Statistical Measures of Forecast Accuracy', *Journal of Forecasting*, **19**, 537–560 (2000).

Greer, M.R., 'Directional Accuracy Tests of Long–Term Interest Rate Forecasts', *International Journal of Forecasting*, **19**, 291–298 (2003).

Greer, M.R., 'Combination Forecasting for Directional Accuracy: An Application to Survey Interest Rate Forecasts', *Journal of Applied Statistics*, **32**, 607–615 (2005).

Hall P., 'Resampling a Coverage Process' *Stochastic Process Applications*, **19**, 259–269 (1985).

Hall P. and J.L. Horowitz, 'Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators' *Econometrica*, **64**, 891–916 (1996).

Hall P., J.L. Horowitz and B. Jing, 'On Blocking Rules for the Bootstrap with Dependent Data' *Biometrika*, **82**, 561–574 (1995).

Havenner, A. and B. Modjtahedi, 'Foreign Exchange Rates: A Multiple Currency and Maturity Analysis' *Journal of Econometrics*, **37**, 251–264 (1988).

Henriksson, R.D. and R.C. Merton, 'On Market Timing and Investment Performance II: Statistical Procedures for Evaluating Forecasting Skills', *Journal of Business*, **54**, 513–533 (1981).

Herwartz, H. and L. Morales–Arias, 'In–Sample and Out–of–Sample Properties of International Stock Return Dynamics Conditional on Equilibrium Pricing Factors' *Christian–Albrechts–Universität zu Kiel, Economics Working Paper* (2008).

Holt, D., A.J. Scott and P.D. Ewings, 'Chi–Squared Tests with Survey Data' *Journal of the Royal Statistical Society*, **Ser. A, 143**, 303–320 (1980).

Inoue A. and M. Shintani, 'Bootstrapping GMM Estimators for Time Series' *Journal of Econometrics*, **133**, 531–555 (2006).

Joutz, F. and H.O. Stekler, 'Data Revisions and Forecasting', *Applied Economics*, **30**, 1011–1016 (1998).

Joutz, F. and H.O. Stekler, 'An Evaluation of the Predictions of the Federal Reserve', *International Journal of Forecasting*, **16**, 17–38 (2000).

Kolb, R.A. and H.O. Stekler, 'How Well Do Analysts Forecast Interest Rates', *Journal of Forecasting*, **15**, 385–394 (1996).

Kuan, C.M. and T. Liu, 'Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks', *Journal of Applied Econometrics*, **10**, 347–364 (1995).

Künsch, H.R., 'The Jackknife and the Bootstrap for General Stationary Observations', *Annals of Statistics*, **17**, 1217–1241 (1989).

Lai K.S., 'An Evaluation of Survey Exchange Rate Forecasts', *Economics Letters*, **32**, 61–65 (1990).

Lahiri, S.N., 'Second Order Optimality of Stationary Bootstrap', *Statistics & Probability Letters*, **11**, 335–341 (1991).

Lahiri, S.N., 'Edgeworth Correction by Moving Block Bootstrap for Stationary and Nonstationary Data', *In LePage R. and L. Billard (eds.), Exploring the Limits of Bootstrap*, Wiley, New York (1992).

Lahiri, S.N., 'On Edgeworth Expansion and Moving Block Bootstrap for Studentized M–Estimators in Multipe Linear Regression Models', *Journal of Multivariate Analysis*, **56**, 42–59 (1996).

Lahiri, S.N., 'Theoretical Comparisons of Block Bootstrap Methods', *Annals of Statistics*, **27**, 386–404 (1999).

Lahiri, S.N., K. Furukawa and Y.D. Lee, 'A Nonparametric Plug–In Rule for Selecting Optimal Block Lengths for Block Bootstrap Methods', *Statistical Methodology*, **4**, 292–321 (2007).

Leitch G. and J.E. Tanner, 'Professional Economic Forecasts: Are They Worth Their Costs?', *Journal of Forecasting*, **14**, 143–157 (1995).

Liu, R.Y. and K. Singh, 'Moving Blocks Jackknife and Bootstrap Capture Weak Dependence', *In LePage R. and L. Billard (eds.), Exploring the Limits of Bootstrap*, Wiley, New York (1992).

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

Merton, R.C., 'On Market Timing and Investment Performance I: An Equilibrium Theory of Value for Market Forecasts', *Journal of Business*, **54**, 363–406 (1981).

Mills, C.T. and G.T. Pepper, 'Assessing the Forecasters: An Analysis of the Forecasting Records of the Treasury, the London Business School and the National Institute', *International Journal of Forecasting*, **15**, 247–257 (1999).

Newey, W.K. and K.D. West, 'A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica*, **55**, 703-708 (1987).

Newey, W.K. and K.D. West, 'Automatic lag selection in covariance matrix estimation', *Review of Economic Studies*, **61**, 631–653 (1994).

Öller L.E. and B. Barot, 'The Accuracy of European Growth and Inflation Forecasts', *International Journal of Forecasting*, **16**, 293–315 (2000).

Patankar, V.N, 'The Goodness of Fit of Frequency Distributions Obtained from Stochastic Processes', *Biometrika*, **41**, 450–462 (1954).

Pesaran, M.H. and S. Skouras, 'Decision–Based Methods for Forecast Evaluation', *In Clements, M.P., and D.F. Hendry (eds.), A Companion to Economic Forecasting*, Oxford, Blackwell Publishing (2002).

Pesaran, M.H. and A.G. Timmerman, 'A Simple Nonparametric Test of Predictive Performance', *Journal of Business & Economic Statistics*, **10**, 461–465 (1992).

Pesaran, M.H. and A.G. Timmerman, 'Predictability of Stock Returns: Robustness and Economic Significance', *Journal of Finance*, **50**, 1201–1228 (1995).

Pesaran, M.H. and A.G. Timmerman, 'Testing Dependence among Serially Correlated Multi–Category Variables', forthcoming in the *Journal of American Statistical Association* (2008).

Politis, D.N. and J.P. Romano, 'A Circular Block–Resampling Procedure for Stationary Data', *In LePage R. and L. Billard (eds.), Exploring the Limits of Bootstrap*, Wiley, New York (1992).

Politis, D.N. and J.P. Romano, 'The Stationary Bootstrap', *Journal of the American Statistical Association*, **89**, 1303–1313 (1994).

Politis, D.N. and H. White, 'Automatic Block–Length Selection for the Dependent Bootstrap', *Econometric Reviews*, **23**, 53–70 (2004).

Pons, J., 'The Accuracy of IMF and OECD Forecasts for G7 Countries', *Journal of Forecasting*, **19**, 53–63 (2000).

Pons, J., 'The Rationality of Price Forecasts: A Directional Analysis', *Applied Financial Economics*, **11**, 287–290 (2001).

Rao, J.N.K. and A.J. Scott, 'The Analysis of Categorical Data From Complex Sample Surveys: Chi–Squared Tests for Goodness of Fit and Independence in Two-Way Tables', *Journal of the American Statistical Association*, **76**, 221–230 (1981).

Radulovic D., 'The Bootstrap of the Mean for Strong Mixing Sequences Under Minimal Conditions', *Statistics & Probability Letters*, **28**, 65–72 (1996).

Schnader, M.H. and H.O. Stekler, 'Evaluating Predictions of Change', *Journal of Business*, **63**, 99–107 (1990).

Schneider, M. and M. Spitzer, 'Forecasting Austrian GDP Using the Generalized Dynamic Factor Model', *Oesterreichische Nationalbank (Austrian Central Bank), Working Paper No. 89* (2004).

Stekler, H.O., 'Are Economic Forecasts Valuable?', *Journal of Forecasting*, **13**, 495–505 (1994).

Stekler, H.O. and G. Petrei, 'Diagnostics for Evaluating the Value and Rationality of Economic Forecasts', *International Journal of Forecasting*, **19**, 735–742 (2003).

Swanson, N.R. and H. White, 'A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks', *Journal of Business & Economic Statistics*, **13**, 265–275, (1995).

Swanson, N.R. and H. White, 'A Model Selection Approach to Real–Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks', *Review of Economics and Business Statistics*, **79**, 540–550, (1997a).

Swanson, N.R. and H. White, 'Forecasting Economic Time Series Using Flexible Versus Fixed Specification and Linear Versus Non Linear Econometric Models', *International Journal of Forecasting*, **13**, 439–461, (1997b).

Tavaré, S. and P.M.E. Altham, 'Dependence in Goodness of Fit Tests and Contingency Tables', *Biometrika*, **70**, 139–144 (1983).

West, K.D., 'Forecast Evaluation', *In Elliot, G., W.J. Granger and A. Timmermann. (eds.), Handbook of Economic Forecasting*, Elsevier B.V. (2006).

Zvingelis, J.J., 'On Bootstrap Coverage Probability With Dependent Data', *In D. Gilles (ed.), Computer–Aided Econometrics*, New York: Marcel Dekker (2001).

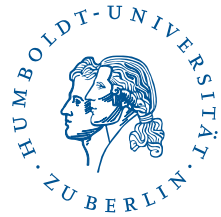
SFB 649 Discussion Paper Series 2008

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Testing Monotonicity of Pricing Kernels" by Yuri Golubev, Wolfgang Härdle and Roman Timonfeev, January 2008.
002 "Adaptive pointwise estimation in time-inhomogeneous time-series models" by Pavel Cizek, Wolfgang Härdle and Vladimir Spokoiny, January 2008.
003 "The Bayesian Additive Classification Tree Applied to Credit Risk Modelling" by Junni L. Zhang and Wolfgang Härdle, January 2008.
004 "Independent Component Analysis Via Copula Techniques" by Ray-Bing Chen, Meihui Guo, Wolfgang Härdle and Shih-Feng Huang, January 2008.
005 "The Default Risk of Firms Examined with Smooth Support Vector Machines" by Wolfgang Härdle, Yuh-Jye Lee, Dorothea Schäfer and Yi-Ren Yeh, January 2008.
006 "Value-at-Risk and Expected Shortfall when there is long range dependence" by Wolfgang Härdle and Julius Mungo, Januray 2008.
007 "A Consistent Nonparametric Test for Causality in Quantile" by Kiho Jeong and Wolfgang Härdle, January 2008.
008 "Do Legal Standards Affect Ethical Concerns of Consumers?" by Dirk Engelmann and Dorothea Kübler, January 2008.
009 "Recursive Portfolio Selection with Decision Trees" by Anton Andriyashin, Wolfgang Härdle and Roman Timofeev, January 2008.
010 "Do Public Banks have a Competitive Advantage?" by Astrid Matthey, January 2008.
011 "Don't aim too high: the potential costs of high aspirations" by Astrid Matthey and Nadja Dwenger, January 2008.
012 "Visualizing exploratory factor analysis models" by Sigbert Klinke and Cornelia Wagner, January 2008.
013 "House Prices and Replacement Cost: A Micro-Level Analysis" by Rainer Schulz and Axel Werwatz, January 2008.
014 "Support Vector Regression Based GARCH Model with Application to Forecasting Volatility of Financial Returns" by Shiyi Chen, Kiho Jeong and Wolfgang Härdle, January 2008.
015 "Structural Constant Conditional Correlation" by Enzo Weber, January 2008.
016 "Estimating Investment Equations in Imperfect Capital Markets" by Silke Hüttel, Oliver Mußhoff, Martin Odening and Nataliya Zinych, January 2008.
017 "Adaptive Forecasting of the EURIBOR Swap Term Structure" by Oliver Blaskowitz and Helmut Herwatz, January 2008.
018 "Solving, Estimating and Selecting Nonlinear Dynamic Models without the Curse of Dimensionality" by Viktor Winschel and Markus Krätzig, February 2008.
019 "The Accuracy of Long-term Real Estate Valuations" by Rainer Schulz, Markus Staiber, Martin Wersing and Axel Werwatz, February 2008.
020 "The Impact of International Outsourcing on Labour Market Dynamics in Germany" by Ronald Bachmann and Sebastian Braun, February 2008.
021 "Preferences for Collective versus Individualised Wage Setting" by Tito Boeri and Michael C. Burda, February 2008.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

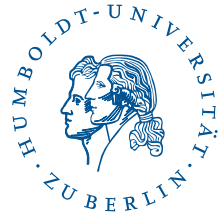


022 "Lumpy Labor Adjustment as a Propagation Mechanism of Business Cycles" by Fang Yao, February 2008.
023 "Family Management, Family Ownership and Downsizing: Evidence from S&P 500 Firms" by Jörn Hendrich Block, February 2008.
024 "Skill Specific Unemployment with Imperfect Substitution of Skills" by Runli Xie, March 2008.
025 "Price Adjustment to News with Uncertain Precision" by Nikolaus Hautsch, Dieter Hess and Christoph Müller, March 2008.
026 "Information and Beliefs in a Repeated Normal-form Game" by Dietmar Fehr, Dorothea Kübler and David Danz, March 2008.
027 "The Stochastic Fluctuation of the Quantile Regression Curve" by Wolfgang Härdle and Song Song, March 2008.
028 "Are stewardship and valuation usefulness compatible or alternative objectives of financial accounting?" by Joachim Gassen, March 2008.
029 "Genetic Codes of Mergers, Post Merger Technology Evolution and Why Mergers Fail" by Alexander Cuntz, April 2008.
030 "Using R, LaTeX and Wiki for an Arabic e-learning platform" by Taleb Ahmad, Wolfgang Härdle, Sigbert Klinke and Shafeeqah Al Awadhi, April 2008.
031 "Beyond the business cycle – factors driving aggregate mortality rates" by Katja Hanewald, April 2008.
032 "Against All Odds? National Sentiment and Wagering on European Football" by Sebastian Braun and Michael Kvasnicka, April 2008.
033 "Are CEOs in Family Firms Paid Like Bureaucrats? Evidence from Bayesian and Frequentist Analyses" by Jörn Hendrich Block, April 2008.
034 "JBendge: An Object-Oriented System for Solving, Estimating and Selecting Nonlinear Dynamic Models" by Viktor Winschel and Markus Krätzig, April 2008.
035 "Stock Picking via Nonsymmetrically Pruned Binary Decision Trees" by Anton Andriyashin, May 2008.
036 "Expected Inflation, Expected Stock Returns, and Money Illusion: What can we learn from Survey Expectations?" by Maik Schmeling and Andreas Schrimpf, May 2008.
037 "The Impact of Individual Investment Behavior for Retirement Welfare: Evidence from the United States and Germany" by Thomas Post, Helmut Gründl, Joan T. Schmit and Anja Zimmer, May 2008.
038 "Dynamic Semiparametric Factor Models in Risk Neutral Density Estimation" by Enzo Giacomini, Wolfgang Härdle and Volker Krätschmer, May 2008.
039 "Can Education Save Europe From High Unemployment?" by Nicole Walter and Runli Xie, June 2008.
040 "Solow Residuals without Capital Stocks" by Michael C. Burda and Battista Severgnini, August 2008.
041 "Unionization, Stochastic Dominance, and Compression of the Wage Distribution: Evidence from Germany" by Michael C. Burda, Bernd Fitzenberger, Alexander Lembcke and Thorsten Vogel, March 2008
042 "Gruppenvergleiche bei hypothetischen Konstrukten – Die Prüfung der Übereinstimmung von Messmodellen mit der Strukturgleichungs-methodik" by Dirk Temme and Lutz Hildebrandt, June 2008.
043 "Modeling Dependencies in Finance using Copulae" by Wolfgang Härdle, Ostap Okhrin and Yarema Okhrin, June 2008.
044 "Numerics of Implied Binomial Trees" by Wolfgang Härdle and Alena Mysickova, June 2008.

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

045 "Measuring and Modeling Risk Using High-Frequency Data" by Wolfgang Härdle, Nikolaus Hautsch and Uta Pigorsch, June 2008.
046 "Links between sustainability-related innovation and sustainability management" by Marcus Wagner, June 2008.
047 "Modelling High-Frequency Volatility and Liquidity Using Multiplicative Error Models" by Nikolaus Hautsch and Vahidin Jeleskovic, July 2008.
048 "Macro Wine in Financial Skins: The Oil-FX Interdependence" by Enzo Weber, July 2008.
049 "Simultaneous Stochastic Volatility Transmission Across American Equity Markets" by Enzo Weber, July 2008.
050 "A semiparametric factor model for electricity forward curve dynamics" by Szymon Borak and Rafał Weron, July 2008.
051 "Recurrent Support Vector Regreson for a Nonlinear ARMA Model with Applications to Forecasting Financial Returns" by Shiyi Chen, Kiho Jeong and Wolfgang K. Härdle, July 2008.
052 "Bayesian Demographic Modeling and Forecasting: An Application to U.S. Mortality" by Wolfgang Reichmuth and Samad Sarferaz, July 2008.
053 "Yield Curve Factors, Term Structure Volatility, and Bond Risk Premia" by Nikolaus Hautsch and Yangguoyi Ou, July 2008.
054 "The Natural Rate Hypothesis and Real Determinacy" by Alexander Meyer-Gohde, July 2008.
055 "Technology sourcing by large incumbents through acquisition of small firms" by Marcus Wagner, July 2008.
056 "Lumpy Labor Adjustment as a Propagation Mechanism of Business Cycle" by Fang Yao, August 2008.
057 "Measuring changes in preferences and perception due to the entry of a new brand with choice data" by Lutz Hildebrandt and Lea Kalweit, August 2008.
058 "Statistics E-learning Platforms: Evaluation Case Studies" by Taleb Ahmad and Wolfgang Härdle, August 2008.
059 "The Influence of the Business Cycle on Mortality" by Wolfgang H. Reichmuth and Samad Sarferaz, September 2008.
060 "Matching Theory and Data: Bayesian Vector Autoregression and Dynamic Stochastic General Equilibrium Models" by Alexander Kriwoluzky, September 2008.
061 "Eine Analyse der Dimensionen des Fortune-Reputationsindex" by Lutz Hildebrandt, Henning Kreis and Joachim Schwalbach, September 2008.
062 "Nonlinear Modeling of Target Leverage with Latent Determinant Variables – New Evidence on the Trade-off Theory" by Ralf Sabiwalsky, September 2008.
063 "Discrete-Time Stochastic Volatility Models and MCMC-Based Statistical Inference" by Nikolaus Hautsch and Yangguoyi Ou, September 2008.
064 "A note on the model selection risk for ANOVA based adaptive forecasting of the EURIBOR swap term structure" by Oliver Blaskowitz and Helmut Herwartz, October 2008.
065 "When, How Fast and by How Much do Trade Costs change in the EURO Area?" by Helmut Herwartz and Henning Weber, October 2008.
066 "The U.S. Business Cycle, 1867-1995: Dynamic Factor Analysis vs. Reconstructed National Accounts" by Albrecht Ritschl, Samad Sarferaz and Martin Uebele, November 2008.
067 "Testing Multiplicative Error Models Using Conditional Moment Tests" by Nikolaus Hautsch, November 2008.
068 "Understanding West German Economic Growth in the 1950s" by Barry Eichengreen and Albrecht Ritschl, December 2008.

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".



069 "Structural Dynamic Conditional Correlation" by Enzo Weber, December 2008.

070 "A Brand Specific Investigation of International Cost Shock Threats on Price and Margin with a Manufacturer-Wholesaler-Retailer Model" by Till Dannewald and Lutz Hildebrandt, December 2008.

071 "Winners and Losers of Early Elections: On the Welfare Implications of Political Blockades and Early Elections" by Felix Bierbrauer and Lydia Mechtenberg, December 2008.

072 "Common Influences, Spillover and Integration in Chinese Stock Markets" by Enzo Weber and Yanqun Zhang, December 2008.

073 "Testing directional forecast value in the presence of serial correlation" by Oliver Blaskowitz and Helmut Herwartz, December 2008.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".